An Item Bank for Testing English Language Proficiency

using the Rasch model to construct an objective measure

by

Neil Jones

PhD

University of Edinburgh

1992

For Iwona, Alexander and Zosia

Declaration

I declare that this thesis has been composed by myself and that the work involved is entirely my own.

Acknowledgements

This study started with a *tachi-yomi* or standing-up read in a bookshop in London, when I happened to pick up *A Guide to LanguageTesting* by Grant Henning, and found the two chapters on Latent Trait Theory. Back in Poland I went about furiously constructing traits. So I should begin by thanking the participants of the English courses at the Adam Mickiewicz University, Poznan, who bore with patience the barrage of trial tests inflicted on them over two years.

Sincere thanks are due to Brian North, whose interest in this early item banking activity convinced me that it was worth pursuing more seriously, and who has been instrumental in co-ordinating the trialling of items in Eurocentres schools, and moving the ItemBanker software project forwards. Thanks too to all the teachers at Eurocentres who administered and marked trial tests, and provided a great deal of useful comment: particularly to Rosey Feuell, SueEvans, David King, Ian Anderson and Martyn Ellis. I gratefully acknowledge the generous financial support offered by Eurocentres in the second and third years of this study: without this it would never have got beyond Year One.

I would like to thank Clive Criper, my supervisor in my first year at Edinburgh; Alastair Pollitt, who agreed to take on a supervisory role from the remoteness of Cambridge, and who has given me invaluable help with understanding and using the Rasch model; and Alan Davies, who returned to Edinburgh to provide supervision and encouragement in my third year, when I needed it most.

Thanks too to Ben Wright and all participants of the 5th ObjectiveMeasurement Workshop in Chicago, who heard out a presentation without a single number in it, and were generous enough about it.

Gwyneth Fox of Cobuild kindly provided a useful word-frequency list.

Above all, of course, thanks to Iwona, for abandoning the Iron Curtain (it was still there then) to live in Edinburgh, and to our children Alexander and Zosia, who put the importance of the whole enterprise in proper perspective.

Abstract

This study describes the construction of an instrument for testing English language proficiency: a bank of about a thousand quite heterogeneous items, covering a range from beginner level to advanced. The software is specially written, to enable teachers to make tests easily, choosing level and content areas; it also supports computer-adaptive testing, with more task variety than has been usual.

The Rasch item response model is used to locate the items on a single difficulty scale. Rasch analysis makes possible the *objective measurement* of psychological traits, which means essentially that constructs having no physical counterpart, like language proficiency, can be treated analogously to physical objects, quantities of which can be measured in conventional fixed units.

The question is asked whether language proficiency *can* be conceived of in simple enough terms to make objective measurement feasible. A review of the fields of second-language acquisition studies, language testing and teaching concludes that language proficiency (in some aspect) is a reasonable candidate for the construction of a unidimensional trait. Analysis of the items in the bank confirms that they fit to a unidimensional trait, and that the Rasch model performs satisfactorily, although calibrations of badly-targetted items are distorted. A multiple regression analysis is used to investigate item difficulty, and thus what it is that the bank really measures. A causal model in which an item's *content* (the language problem tested) is placed first finds *method facets* (e.g. the form of response) to be weak predictors of difficulty. What makes language test items difficult, it is concluded, is mostly the difficulty of the language problems tested. Qualitative analysis of items grouped by content is also informative. It appears that item difficulty is largely (though not entirely) explicable in terms of factors that should be included in a theory of language learning.

Table of Contents

Chapter 1: Introduction *1*

Chapter 2: Language Proficiency *8*

2.1 Language ability 9
2.1.1 Interlanguage 9
2.1.2 Competence and performance 12
2.1.3 The role of conscious knowledge 16
2.1.4 The testing viewpoint: the Unitary Competence Hypothesis 18
2.1.5 Communicative competence 22
2.2 Language difficulty; the development of proficiency 26
2.2.1 The natural order hypothesis 27
2.2.2 Does teaching make a difference? 29
2.2.3 A grammatical view of development 31
2.2.4 Linguistic universals 34
2.2.5 Variability and first language transfer 36
2.2.6 The teaching viewpoint: organisation and grading 38
2.2.7 Language development and cognitive difficulty 42
2.3 Discussion 46
2.3.1 General language proficiency 46
2.3.2 Conclusion 56

Chapter 3: Item Response Theory *61*

3.1 The construct of language proficiency 61
3.2 An introduction to Item Response Theory 62
3.2.1 Shortcomings of standard testing methods 62
3.2.2 Item Response Theory 63
3.2.3 Advantages claimed for IRT 68
3.2.4 Assumptions of IRT 70

3.2.5 Choosing an IRT model 71
3.3 Construct validation and IRT 75
3.3.1 Construct validation 75
3.3.2 Unidimensionality 77
3.3.3 Explaining item difficulty 85

Chapter 4: The design of the item bank *90*

4.1 Item Banker in outline 90
4.2 Simple operation by teachers 90
4.2.1 Overview screen 91
4.2.2 Search screen 94
4.2.3 Item viewing screen 97
4.2.4 Printing screen 98
4.3 Operation by 'system users' 98
4.3.1 Writing and editing items 99
4.3.2 Calibration of new items 101
4.3.3 Other facilities 103
4.4 Printed output from ItemBanker 103
4.5 The Computer-adaptive test 104
4.6 Appendix: A sample Item Banker test 108

Chapter 5: Data collection, item calibration and model fit *118*

5.1 Data collection 118
5.1.1 Writing items 118
5.1.2 Linking of items in test forms 122
5.1.3 Feedback during trialling, and revisions made 122
5.1.4 Checking marking 123
5.1.5 Retrialling 124
5.2 Item calibration 125
5.2.1 Approaches to estimation 125
5.2.2 Spread of difficulties 127
5.2.3 Poorly estimated items 127
5.3 Investigating model-data fit 131
5.3.1 Testing model assumptions 131
5.3.2 Checking model features (Invariance of estimates) 133

5.3.3 Investigating model fit 141
5.3.3.1 Alternative marking schemes and fit 143
5.3.3.2 Item types and fit 144
5.3.3.3 Difficulty and fit 146
5.3.3.4 Item content and fit 147
5.3.3.5 The influence of L1 on fit 148
5.4 Appendix: A note on Rasch estimation of item difficulties 152

Chapter 6: Explaining item difficulty *158*

6.1 Causal models, levels of abstraction 158
6.2 A model for test item difficulty 161
6.2.1 The language problem is logically prior 162
6.2.2 Rubric 165
6.2.3 The Prompt 166
6.2.4 The Response 167
6.2.5 Marking the paper 170
6.3 A Multiple-Regression analysis 170
6.4 A qualitative analysis of groups of items 177
6.4.1 The 'Language Problem' defined 179
6.4.2 Examples from the bank 182
6.5 Appendix: Examples of language problems 193

Chapter 7: Conclusion *201*

7.1 Rasch analysis and vertical equating 201
7.2 Constructing the trait: model fit 203
7.3 Interpreting the trait 204
7.4 Uses of the item bank 208

References 212

1: Introduction

One speaks of language proficiency testing as a form of *measurement*; indeed, the word *psychometrics*, denoting the larger field of study of which language testing is a part, suggests no less. But in fact, a look in the Oxford English Dictionary should convince us that language testing is not measurement at all - at least, not in the strictest sense. The verb *measure* is defined as follows:

> To ascertain or determine the spatial magnitude or quantity of (something); *properly*, by the application of some object of known size or capacity. Also, in extended sense, to ascertain the quantity of (e.g. force, heat, time) by comparison with some fixed unit (OED, italics in original).

The reason that most language testing does not qualify as measurement in the strict sense is the lack of any 'comparison with some fixed unit'. The comparison through which scores in a test become interpretable is generally with the performance of other people taking the same test, or sometimes with criterion tasks that indicate 'mastery' of the field tested.

This is not to suggest that the whole language testing enterprise to date is in some way flawed, but simply to point out the power of metaphor to colour our understanding of what it is that we are doing. We tend to think of scores in language tests as if they represented simple attributes, like length or weight, which have the property of *additivity* essential to true measurement; but the fact is that hitherto we have had no easy way of checking whether this is, or can be, the case.

The desirability of having 'fixed units' with which to describe language proficiency has in recent years led to several attempts to construct 'scales' or 'frameworks'. The ESU Framework (Carroll & West, 1989), for example, attempts to compare the various examinations of the main English Language Boards, using

'a model which describes each examination in a standard way, drawing on a series of performance scales or "Yardsticks"' (Carroll & West, 1989:2). It is interesting that the choice of the term *yardstick* appeals directly to the spatial measurement metaphor. The meaning of such scales is derived from a set of 'band descriptors', which describe the kind of language performance which is characteristic of each level. For example, the 'Linguistic skills' yardstick of the ESU Framework includes the following band descriptor:

> Applies linguistic skills to moderate-level tasks with adequate confidence and competence. Presentation of basic message is adequately adjusted to audience's knowledge of the language. Fairly frequent language lapses necessitate repair to capture detail and sublety. Basic organisation of text is adequate, with a moderate range of cohesive devices.... (Carroll & West, 1989:58)

It is striking that such descriptions taken on their own are inadequate to define a level. Nothing about this text, for example, would lead us to the conclusion that it describes band five of a nine-band scale (rather than, say, band six, four, or three). While terms such as *moderate*, *adequate*, *fairly* take on more meaning in the context of the whole set of descriptors, it is clear that the accurate application of such a scale relies upon judgement, experience, and comparison with samples of criterion performance for each level.

Again, this is not to suggest that such scales are not useful, or that assessments based on the exercise of judgement are inherently unsatisfactory (they are not). But it should be clear that it is the *appearance* of true measurement which is being achieved here. The very generality of the description may be a virtue for some purposes, but a detailed and particular characterisation, stating what parts of the language system a

learner would typically control at each level, would be much more useful. A reviewer of language testing over the last ten years states that

> One of the most serious shortcomings ... has been the lack of progress in (or even concern for) the *growth* and *development* in proficiency. (Skehan 1989:9)

If this is so, then a major cause must be the lack of a suitable measuring instrument.

The present study addresses this issue. It describes the construction of a new testing instrument: an item bank comprising a collection of about one thousand English language test items, held in a specially-written computer database. The difficulty of each item has been found (using quantitative methods), and is expressed in terms of a single scale which covers a proficiency range from beginner level to very advanced. A learner's proficiency level, as measured by items from the bank, can be reported in terms of this single scale.

The benefits are considerable. Learners can be placed in a teaching progamme, or their progress measured, entirely objectively (in the sense that no great exercise of judgement is necessary). Measures are also objective in the sense that they are not relative to any particular group of persons. The coherence and uniformity of the language proficiency trait depicted by the items in the bank is not taken on trust, but can be investigated. Perhaps most interestingly, the items themselves constitute a detailed description of the developing language proficiency trait, and of the knowledge that learners typically have at different levels. The item bank may thus prove to have applications not only for placement and progress testing, but for diagnostic testing, syllabus design, and possibly even in second language acquisition research.

This study covers three major areas: the design of the bank itself, the construction and trialling of items to go in the bank, and the investigation of the nature of the proficiency trait which is depicted by the items, once their difficulty has been found. The first two areas may be seen as the more practical part of the study, while the third provides the theoretical focus. The following section reviews some of the issues that will be dealt with.

Overview

Firstly, it will be necessary to investigate the notion of language proficiency, in order to arrive at a working definition to inform the selection of items for the bank. Language proficiency has been characterised as a single, unanalysable attribute which different people possess in differing quantities, or alternatively as a bundle of loosely-related skills or competences, each of which can be separately measured. There is of course truth in both of these views, and both have at different times been espoused by language testers. Which view one adopts must depend largely on the context of the test: the reason for testing, the kind of inferences to be drawn from learners' performance. The present study is committed *a priori* to defining language proficiency in terms that allow it to be treated as a unitary construct, and does so simply because it *must* be so treated if we are to construct an instrument to measure it (just as a ruler can measure only length, and a pair of scales only weight). This requirement of *unidimensionality* will be introduced in Chapter 2 and returned to at greater length in the presentation of Item Response Theory in Chapter 3.

The narrower the definition of language proficiency, the better the chances of constructing an instrument to measure it. The discussion in Chapter 2 will lead us to identify a 'core' aspect of language proficiency: the sort of area characterised as

'linguistic competence' or 'organisational competence' in taxonomies of language ability. In other words, grammatical knowledge is considered a central aspect of language proficiency.

This choice of focus follows from the item bank's intended role as an instrument for *formative* assessment within a teaching programme, rather than *summative* assessment at the end of a course. The items in the bank are short, discrete items testing chiefly knowledge of vocabulary and the language system. Because many items have a clear pedagogic point, performance on a test can provide a detailed recipe for remedial action on the part of individual learners. The item bank thus exemplifies an 'indirect' approach to testing language proficiency, as opposed to more 'direct' tests of communicative language ability. Item bank tests are unspeeded, lack 'authentic' communicative purpose, and allow the testee recourse to explicit, conscious knowledge. This (it can be argued) makes them very poor candidates for constructing a language proficiency trait which is interpretable in 'developmental' terms. This issue is considered in Chapter 2, and taken up again in the final discussion (Chapter 7), where it is concluded that the trait is both constructable and interpretable. More is said about the selection of items for the bank in Chapter 5.

Chapter 3 introduces Item Response Theory (IRT) - the statistical theory upon which the item bank depends - and draws together the notions of language ability and item difficulty. IRT relates the probability of a person answering an item correctly to the difference between the ability of the person and the difficulty of the item. IRT provides methods for estimating difficulties and abilities, given an actual set of test scores. Furthermore, because the difficulty/ability scale is linear (like a ruler) it should be a relatively simple matter in IRT to link results from several different test administrations to the same scale. Thus using IRT it is feasible to extend the scale to cover the whole range of ability we are interested in testing.

Chapter 4 describes the design of the present item bank. The chief considerations in designing the bank were that it should be easy to use by teachers to construct tests at desired proficiency levels and on chosen content areas; and that it should also support computer-adaptive testing, while offering a wider variety of task type than has been customary in computer-adaptive testing to date.

Chapter 5 describes experimental work: the construction of the set of items for the bank, and their trialling. Practical problems with using Rasch estimation (the branch of IRT used in the present study) are discussed; in particular it is found that problems arise when items are badly targetted - that is, trialled on learners of inappropriate level. The findings discussed here may thus be seen to contribute to the debate over the suitability of the Rasch model for vertical equating. The conclusion is that Rasch estimation *can* satisfactorily locate items of widely-differing difficulty on a single scale, although care must be taken with the choice of data.

Chapter 5 also examines whether the items do in fact demonstrate invariable difficulties, and the extent to which, taken together, they delineate a single, coherent trait. This represents the first stage of construct validation: that is, the demonstration that the construct which the bank purports to measure does in fact have some coherent shape.

Chapter 6 pursues construct validation further by attempting to explain what it is that makes items difficult, and thus what it is that the bank really measures. Multiple-regression analysis is used in an attempt to quantify the contribution to difficulty made by a range of item features. The important point is made that in order to *explain* item difficulty (rather than simply *describe* the features of difficult and easy items) it is necessary to propose a *causal* theory, which in multiple-regression terms means specifying (and justifying) the order in which item features are added to the equation. It is argued that

the *content* of items (what they test) must be treated as logically prior to *method* factors (how they test it). In the case of the present item bank, it is the *language problem* which is logically prior. The multiple-regression analysis shows that under this causal theory, method factors, such as the form of the response, the number of words to be written, even the difficulty of the vocabulary used, cease to be strong predictors of item difficulty. What makes language test items difficult is, first and foremost, the difficulty of the language problems tested.

In the present bank the 'language problems' that can be studied are chiefly traditional pedagogic points, such as 'the First Conditional'. It is possible to use the items to depict the 'difficulty envelopes' of a number of such problems - that is, the range of difficulty which they cover in the bank. The easiest items on a particular problem typically offer a great deal of support to the learner. Then come items embodying formulaic use, or use in familiar contexts. Harder items tend to embody use in more abstract, cognitively-demanding contexts, or sometimes invoke cultural or conventional knowledge. Such qualitative analyses of groups of items are frequently informative, and reinforce the impression that item difficulty is largely (though not entirely) explicable in terms of factors that we would wish to include in a theory of language learning.

2: Language proficiency

It was stated in the introduction that in order to measure objectively a psychological attribute such as language proficiency we have to be able to treat it as if it were some simple attribute like length or weight. Let us briefly consider what this implies, using the analogy with measurement of attributes of physical objects by way of illustration.

Firstly, it must be a *unidimensional* trait. Physical objects have length, weight, and many other attributes, but you can only compare them using *one* attribute at a time; you cannot, for example, meaningfully compare a length of two metres with a weight of five kilograms.

Secondly, it must be invariant, in the sense that a given measuring instrument should remain accurate, whatever particular object you apply it to; and conversely, a given object should be accurately measured, whatever the choice of measuring instrument. If a tape measure provides inaccurate measures of the waistlines of a group of people, this might either be because of a defect in the tape measure (it is elastic, say), or because of variability in the responses of the people (the vainer might try to hold their stomachs in).

Language proficiency, like any psychological attribute, is evidently a very complex phenomenon, yet measurement demands that we *impose* the qualities of unidimensionality and invariance upon it. Thurstone, who pioneered the measurement of psychological attributes, wrote in 1928:

> When we discuss opinions, about prohibition, for example, we quickly find that these opinions are multidimensional, that they cannot all be represented in a linear continuum. The

> various opinions cannot be completely described merely as 'more' or 'less'. They scatter in many dimensions, but the very idea of measurement implies a linear continuum of some sort, such as length, price, volume, weight, age. When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind. And so it is also with attitudes (Thurstone 1959:218).

How this can be done, that is, how a psychological trait can be *constructed*, will be discussed at greater length in the chapter on Item Response Theory. For now we should note that the language proficiency trait involves two complementary notions: language ability, which resides in people, and language difficulty, which resides, let us say, in language itself, or in the world where language is put to use. This chapter is organised around these two notions, looking to three fields - second language acquisition studies, language teaching and language testing.

2.1 Language ability

2.1.1 Interlanguage

The first and central point is that a learner's language ability is not a passive imprint, a more or less imperfect reflection of whatever teaching has come the learner's way. Rather it represents an active effort to impose order and make sense.

It was Corder (1967) who first suggested that learners' errors were not, as behaviourist learning theories maintained, evidence of defective teaching, to be eliminated or avoided if possible. Rather they were significant of a learner's *transitional competence* - an internalized grammar which differed from the

grammar of the target language, but was systematic in its own right. In this he distinguished *errors* from mere *mistakes*, or slips of the tongue:

> The errors of performance will characteristically be unsystematic and the errors of competence systematic. (Corder 1967:166)

Errors were thus evidence of a continuing and creative process of learning, a view which points the similarity between an adult learning a foreign language and the child acquiring its first language. Corder reviews the evidence against this view, including Lenneberg's (1966) findings that physiological changes at puberty make the two processes different; but he maintains that 'it still remains to be shown that the process of learning a second language is of a fundamentally different nature from the process of primary acquisition.' Corder uses the term *built-in syllabus* to describe the process by which the learner revises his transitional competence progressively in the direction of the target language, and speculates that teaching might be more efficient if it could accomodate to this built-in syllabus. Corder does not say whether the built-in syllabus is likely to be universal, or to vary from individual to individual.

Selinker (1969, 1972) introduced the term *interlanguage*, which came to be used in preference to Corder's *transitional competence*. Unlike Corder, he stresses the differences between first language acquisition and second language learning: citing Lenneberg's (1967) concept of the *latent language structure* - an innate mental mechanism that underlies L1 acquisition - he posits the existence of a *latent psychological structure* which can, if called upon, support the learning of a second language. Selinker restricts the notion of interlanguage to the transitional system exhibited by learners exploiting this posited latent psychological structure, on their way towards *fossilization* - that is, a final, stable competence which falls short of the target language. That is, he explicitly excludes from

consideration the small number of learners who finally achieve native-speaker competence, explaining these as rare cases in which the L1 acquisition faculty is somehow re-activated, and speculating that 'these individuals may not go through an IL' (Selinker 1972:223). In this he appears to confuse the contrast between L1 and L2 acquisition, and that between natural acquisition and the explicit knowledge that results from formal instruction:

> This series of assumptions [about the different psychological mechanisms] must be made ... because the second-language learner who actually achieves native-speaker competence cannot possibly have been taught this competence, since linguists are daily ... discovering new and fundamental facts about particular languages. Successful learners ... must have acquired these facts ... *without* having explicitly been taught them.
> (Selinker 1972:213)

If acquisition of the unformulated rules of language is to be the exclusive preserve of a re-activated L1 acquisition faculty, then presumably run-of-the-mill L2 learning consists in assimilating *only* the ready-formulated, teachable facts about language. This is certainly not true.

Selinker (1972) lists five central cognitive processes in the development of interlanguage: *language transfer*; *transfer of training* (i.e. negative effects of teaching); *strategies of second-language learning*; *strategies of second-language communication*; and *overgeneralization* from L2. These are *not* presented as heuristic devices which promote second-language learning, but rather as the source of non-target-language, potentially fossilizing interlanguage features.

> ... each process forces fossilizable material upon surface IL utterances .... Combinations of these processes produce what we might term entirely fossilized IL competences. (Selinker 1972:217)

Adjemian (1976), in contrast to Selinker's cognitive emphasis, and more in line with Corder, proposes a linguistic approach, treating interlanguage as a natural language and attempting to describe the properties of its grammar. In contrast to Selinker's emphasis on fossilization, she stresses the permeable and dynamic nature of interlanguage systems, i.e. their openness to change.

Early interlanguage studies have been characterized as *product* oriented because they took as data descriptions of learner speech, their basic research tool being error analysis. The more recent orientation is towards attempting to explain the underlying *process* of language acquisition, in ways not based wholly on descriptions of speech. These process-oriented approaches draw upon linguistics, sociolinguistics and cognitive science, so that presently 'it is difficult to demarcate where Interlanguage theory ends and other theories begin' (McLaughlin, 1987:80).

2.1.2 Competence and performance

This product-process distinction underlines that language ability has two aspects: knowing *what* - for example, the rules of the grammar - and knowing *how* to make use of language knowledge in order to communicate. Ability is both knowledge and skill, or in other words it involves both *competence* and *performance*.

Chomsky (1965) introduced the distinction between competence and performance. As Campbell & Wales (1970:) point out, he used *performance* in both a weak sense, meaning false starts, deviations from rules, changes of plan in mid-course, and so on,

and a strong sense, denoting the psychological factors involved in perception and production. Performance in the weak sense manifests essentially insignificant features which can be excluded from consideration.

> It is reasonable to assume then that regularities in both the user's knowledge of grammar and knowledge of language use can be abstracted from their actual realization in performance and studied independently of nonessential or non-specific (in Campbell and Wales' 1970 terminology) features of performance.
> (Canale & Swain 1980:6)

In the strong sense, performance is that complex of abilities that make language use possible, and is thus central to SLA theory. *Procedural knowledge*, as performance in this sense can be called, is 'a second kind of competence' (Sorace 1985:239) that links interlanguage knowledge and interlanguage use.

This view has given rise to various dualistic models of second-language acquisition which include both language knowledge and the ability to use that knowledge - simply, knowing *what* and knowing *how* (Jordens 1986, Faerch & Kasper 1986, Bialystok & Sharwood Smith 1985).

Bialystok and Sharwood Smith (1985) propose a dual model taking account of both *competence* and *control*. Control refers to the efficiency of retrieval of linguistic knowledge, that is, it equates with performance in the strong sense of that term described above. A learner's interlanguage reflects what he knows of the language, and how well he can use that knowledge.

Competence comprises representations of both grammatical and pragmatic knowledge. Learners have such knowledge in varying *amounts*, but for Bialystok and Sharwood Smith a more important

factor is the degree of *analysis* applied to this knowledge. By analysis is meant the relation of fragments of language into larger, more abstract systems through the construction of rules.

> The qualitative feature of analysis is more important than the quantitative feature of amount, because it is the former that determines ultimately what the learner will be able to do with the language.
> (Bialystok & Sharwood Smith 1985:107)

Bialystok & Sharwood Smith appear to attach little importance to the distinction between conscious and unconscious knowledge: 'The primary effect of analysis is not to increase the conscious awareness of the system, but to increase the potential for use of that system' (p.107). Arguing from L1 or natural L2 acquisition, (where metalinguistic knowledge necessarily *follows* acquisition) they seem to disregard the possible difference of the L2 formal learning situation, where metalinguistic knowledge may *precede* acquisition.

Control is associated with efficiency of retrieval, automaticity, and is called 'the basis of fluency' (p.109). Further, they claim, 'this fluency is independent of knowledge,' giving as illustration a learner who is able to communicate fluently despite large gaps in his grammatical knowledge. This is rather curious, given their earlier claim that in language learning, development of *representations* of linguistic structure must logically precede the development of *procedures* for retrieving them.

By including performance factors - whether called *control*, *procedural knowledge* or *rules of language production* - in the description of language ability, second-language acquisition studies move closer to general cognitive theory. McLaughlin defines a cognitive process as follows:

> Learning is a *cognitive* process, because it is thought to involve internal representations that regulate and guide performance. In the case of language acquisition, these representations are based on the language system and include procedures for selecting appropriate vocabulary, grammatical rules, and pragmatic conventions governing language use. As performance improves, there is constant restructuring as learners simplify, unify, and gain increasing control over their internal representations.... These two notions - automatization and restructuring - are central to Cognitive theory.
> (McLaughlin 1987:133)

In this view, then, skills are learned and become automatic only after the earlier use of *controlled processes*. Controlled processes regulate the flow of information from short-term to long-term memory. It is the limited capacity of short-term memory that is seen as the barrier to fluency: as certain elements of language use become automatized, they cease to take up short-term memory, and previously difficult tasks become easier. McLaughlin, like Bialystok & Sharwood Smith above, stresses that 'the distinction between controlled and automatic processing is not based on conscious experience. Both controlled and automatic processes can in principle be either conscious or not' (1987:153).

Restructuring - the second central notion in cognitive theory - denotes a qualitative change in the way knowledge is organized - the imposition of a new organizing principle on a variety of hitherto perhaps unrelated or unanalyzed language elements. Restructuring can be seen as driven by *learning strategies* (Ellis 1985). Thus in the early stages a learner may tend to simplify, overgeneralize, and construct a simple picture of the language that relies more on first-language transfer or on language universals. At more advanced levels, strategies may attend more

closely to the second-language data. Restructuring accounts for the quite sudden changes in a learner's performance that have frequently been noted in interlanguage studies.

2.1.3 The role of conscious knowledge

As stated above, the distinction between conscious and unconscious, explicit and implicit knowledge is not particularly strongly drawn in cognitively-based theories of language learning. However, Krashen's well-publicized model of second language acquisition (e.g. Krashen 1981, 1982) makes this a central issue. This is a dual model in which *learning* and *acquisition* are contrasted. Learning is explicit, formal language knowledge characteristic of classrooms, and acquisition is implicit, naturally learned language knowledge characteristic of the way in which children acquire their first language. Acquired knowledge is claimed to be the basis of all spontaneous language use; learned knowledge is unable to contribute to language use except as a *monitor*, in a manner limited by situational constraints (for example, monitoring will be more feasible in slow, careful speech than in spontaneous conversation). The two systems, moreover, are claimed to be unrelated, in the sense that learned knowledge never becomes acquired knowledge, however much drill or practice is applied to it.

The duality proposed by Krashen is not exactly the knowledge/use opposition presented above. Acquisition, in Krashen's sense, seems to encompass both a base of language knowledge (unconscious) *and* the ability to make use of that knowledge in performance. Krashen's distinction is between explicit language knowledge and subconscious learning processes, and his claim is that explicit knowledge is an inessential, secondary aspect of proficiency in a second language.

Krashen's claims have been subjected to much criticism (Brumfit 1984, Gregg 1984, McLaughlin 1987). The learning/acquisition distinction is unverifiable, given the unobservable nature of implicit knowledge. The definitions of learning and acquisition are circular. Faced with a classroom where students' language ability appears to be improving (and there must be such classrooms) the theory can only claim that (useful) acquisition must be taking place somehow alongside (unproductive) learning. Such a claim is unfalsifiable. As to what is meant by a 'learned system', Krashen associates it, as Brumfit (1984:47) points out, with 'conscious, even painstaking application of rules' to the construction of sentences. But conscious study can take many forms, some of which (careful reading, for example) may be useful in a way quite different from the arid memorization of grammar rules. Brumfit summarizes:

> In its strong form, Krashen has the disadvantages of being confused or inexplicit on certain key issues (such as the definition of 'learning'), of being intrinsically unfalsifiable, of conflicting directly with the intuitions of successful language learners and successful language teachers, and of being merely descriptive with no explanatory power.
> (Brumfit 1984:49)

Krashen's learning/acquisition distinction may well have some pedagogic value, but it does not appear to stand up as a theory of learning.

Krashen's *monitor* seems to have something in common with the factor of *attention to speech* which is incorporated by Tarone (1983) or Ellis (1987). Tarone proposes a *capability continuum* of styles, running from the least monitored *vernacular* to the *most careful style*. Each style reflects a different underlying competence, hence the term *heterogeneous competence* used to describe these models. The most careful style is the least stable, because it reflects conscious attention to speech and is

thus more readily influenced, for example, by teaching. Ellis (1987) explains the effect of formal language instruction in this way: newly-learned features will tend to appear first in a learner's careful style, but with practice and increasing automaticity may pass along the capability continuum until they become part of the learner's spontaneous speech. Ellis, then, unlike Krashen, proposes a link between explicit and implicit knowledge.

The heterogeneous competence models owe much to the work of Labov (1966), who showed how native speakers of a language adapt styles of speech to the situation and to their audience. The appropriacy of applying this metaphor of style-shifting to the L2 learner has been questioned. To talk of new language features passing along a capability continuum seems to be a complicated way of saying that practice makes perfect, and indeed, heterogeneous competence models are criticized by Swan (1987) for attaching too much significance to performance factors: 'The data do not add up to anything we could reasonably call a style' (p.62). He observes:

> Notions like *difficulty* and *practice*, which involve considerations of performance, don't seem easy to handle with a model which essentially locates the sources of variability in differential competences.
> (Swan 1987:63)

2.1.4 The testing viewpoint: the Unitary Competence Hypothesis

Second language acquisition studies offer a great many models of language ability, but little evidence. The language testing viewpoint may be more illuminating, at least if we share the view that

> testers, by researching into the structure of language proficiency, are attacking fundamental problems in language teaching and applied linguistics. Testing is where the buck finally stops, and where theorising which does not have empirical consequences should be shown to be vacuous. (Skehan 1989:211)

A scientific approach to language testing can be said to start with what Spolsky (1975) calls the psychometric-structuralist phase. Lado (1961) was the influential theorist of this period, and the most characteristic test type the objectively-marked discrete-point test. This approach reflected the view that knowledge of a language could be assessed as if it were composed of a large number of small elements - *atomistic* in Morrow's (1981:10) term. The aspirations of the approach should not be parodied, however: discrete-point testing soon abandoned the (unrealistic) idea of producing 'pure' items covering only one language point each; and global tests of various kinds were also common. This analytic view of language led naturally to the separate testing of the 'four skills'. The underlying assumption was that language competence is in some way divisible along these lines. The idea of identifying a General Language Proficiency (GLP) underlying performance in the various skills was not so much rejected as neglected. As Vollmer (1981) points out, this *divisible competence* hypothesis was not strongly stated. And there was at the same time the related assumption of *transfer ability* (Davies 1981:185), that is, 'the likelihood of performance on one test being substantially correlated with performance on another.' Or put more ironically, 'it was always recognized that the sum of the whole was greater than any one of the parts' (Davies 1978:216, quoted by Vollmer & Sang 1983:36).

At the end of the sixties Spolsky put the question of language proficiency in a new way: 'What does it mean to know a language or how do you get someone to perform his competence?' (Spolsky 1973). The question heralded a 'one-dimensional' (Vollmer & Sang 1983:36) approach to language ability, the aim being 'to get

beyond the limitation of testing a sample of surface features, and seek rather to tap underlying linguistic competence' (Spolsky 1973:175). It is of course impossible to measure underlying linguistic competence directly - one *can* only look at 'a sample of surface features' and attempt to make inferences from them. Spolsky's statement might thus best be understood to mean that language ability has some central core, and that certain kinds of test might measure this better than others. Spolsky called such tests *competence-oriented*, a term which suggests the aim to 'tap underlying linguistic competence'. Such tests were also called *integrative* (Carroll 1961), which suggests the coming-together in one test of a variety of language skills. The term *pragmatic* was also used (Oller 1978) of integrative tests aimed at testing functional language skills. The test types most researched were dictation and cloze, though reading and listening comprehension, and the oral interview, all qualify as integrative.

The attraction of integrative tests lay not simply in the fact that they corresponded somewhat more closely to normal communicative uses of language, but rather in the statistical finding that of a battery of tests of different skills, all would tend to correlate higher with an integrative test than with each other. Thus the integrative test was seen to be measuring something more central to language proficiency.

The *General Language Proficiency Factor* thus started out as a statistical phenomenon. In Oller's interpretation, however (1974, 1978, 1983), it assumed the status of a psychological entity: a single competence which underlay language performance in all skills. This was the *Unitary Competence Hypothesis*. Oller explained this unitary linguistic competence not as a mere construct, but as a 'real' cognitive mechanism, which he called the *expectancy grammar*. The concept of an expectancy grammar rested on the notion that language understanding worked by analysis-by-synthesis (Neisser 1967); that is, that the listener, working from some expectation of what the speaker is

going to say, attempts to generate a match for what is heard. Productive language use could be viewed in an analogous way: the speaker is guided by some intent to communicate, and continuously monitors his own speech to see if it matches the intended meaning. Productive and receptive language use could thus be treated as different aspects of the same process, in which 'the planning ahead or the hypothesizing about what will come next can be conceptualized in terms of grammar-based expectancies' (Oller 1983:5).

The main statistical evidence which Oller offered in support of the unitary competence hypothesis came from factor analysis, or strictly, principle components analysis. When subtests of a test battery were analyzed, a single factor appeared adequately to account for the variance in scores.

The unitary competence hypothesis generated a considerable amount of research, and was soon subjected to criticism. Most importantly, the factor-analytic techniques used were found to be faulty (Farhady 1983 A, Vollmer & Sang 1983); the very strong first factor was an artifact of the technique, and tended to disappear when the same data were properly analyzed. The psycholinguistic basis for the expectancy grammar was also undermined, as the concept of analysis-by-synthesis was abandoned (Vollmer & Sang 1983:39).

Oller's proposal of a unitary competence to explain performance on different types of test was unnecessary, as Davies (1981:183) points out. An integrative test, by definition, 'contains everything,' and this alone explains its psychometric behaviour. Accordingly the unitary competence hypothesis, having served as a focus for much research in testing, was abandoned in its strongest form, even by its original advocates. The misuse of factor analysis led critics to point out its limitations as a technique for exploring the structure of language proficiency (Vollmer & Sang 1983:70, Davies 1981).

However, Carroll (1983) presents a convincing account of factor analysis. He concludes that, although the factor structure found will vary for any set of tests and learners, proper analysis will probably reveal

> that there are *both* general factors *and* 'divisible' factors of proficiency representing, on the one hand, overall rates of progress in second or foreign language learning, and on the other hand, some specialization of learning rates along such dimensions or aspects of language learning as skill with the spoken language, skill with reading and writing, and skill with pronunciation.
> (Carroll 1983:92)

This reasonable view seems to correspond with what most testing practioners had believed all along.

2.1.5 Communicative competence

In recent years the picture of language ability which testers work with has become hugely more complex, with the rise of *communicative competence* models.

What a learner knows of a language and of how to use it evidently takes in more than grammar. Even taking Chomsky's competence and performance (in the strong sense of the psychological mechanisms employed in language use) together, they exclude (as Hymes 1972, Campbell & Wales 1970 point out) consideration of the *appropriateness* or sociocultural significance of an utterance in context. Communicative competence, as proposed by Hymes, includes not only grammatical competence, but contextual (or sociolinguistic) competence: the ability to use language which is 'not so much *grammatical* but ... *appropriate to the context'* in which it is produced (Campbell & Wales 1970:247), and to the goals that the speaker wishes to achieve.

While the sociolinguistic dimension is evidently an important aspect of much work in second-language acquisition, it is interesting that in McLaughlin's (1987) book-length review of second-language learning the expression 'communicative competence' occurs in only one sentence. The constructs used in the discussion of communicative competence - 'illocutionary competence', 'sociolinguistic competence', etc - appear to be of more relevance to language teachers and testers than to second language acquisition researchers (Hulstijn 1985 B); which is why they have not been introduced earlier.

Canale & Swain (1980) review the theoretical bases of existing 'communicative' approaches to teaching and testing. First they define communicative competence as a broad ability which *includes* grammatical competence and sociolinguistic competence. They argue that communicative *competence* is to be distinguished from communicative *performance* - unlike Hymes, who includes *ability for use* as part of competence. Hymes reasons that noncognitive factors, such as motivation, partly determine competence:

> In speaking of competence, it is especially important not to separate cognitive from affective and volitive factors, so far as the impact of theory on educational practice is concerned.
> (Hymes 1972:283)

They propose a theoretical model in which communicative competence is composed minimally of *grammatical* competence, *sociolinguistic* competence, and *communication strategies*, or *strategic* competence. Canale (1983) adds a fourth component, distinguishing sociolinguistic competence (sociocultural rules) from *discourse* competence (cohesion and coherence). Strategic competence is the ability to improvise or repair breakdowns in communication, and invokes either grammatical or sociolinguistic competence.

Canale & Swain's framework is offered as an aid to balanced syllabus design and development of methodology. In Cziko's terms (1984) it appears to be a *descriptive* rather than a *working model* - that is, it does not explain how these proposed competences exist or relate to each other. But various attempts have been made to validate experimentally this and subsequent frameworks of communicative competence, and the results have been rather mixed.

Allen *et al.* (1983) developed measures of grammatical competence (morphology and syntax), discourse competence (cohesion and coherence) and sociolinguistic competence (sensitivity to register). Factor analysis failed to show that these competences were in fact distinct. Bachman and Palmer (1982) found some evidence for the distinctness of an ability they called 'communicative proficiency'. Their test battery included tests of grammatical competence, pragmatic competence (vocabulary, cohesion and organization) and sociolinguistic competence (sensitivity to register, naturalness, cultural references). Grammatical and pragmatic competence appeared from analysis to be closely related, while sociolinguistic competence appeared to be distinct.

Canale & Swain's model of communicative competence is extended by Bachman (1990) in a model that incorporates *competences, skill factors* and *method factors*. Bachman argues that language tests have a unique feature: that language is both the instrument and the object of measurement.

> What I believe this means is that many characteristics of the instrument, or the method of observing and measuring, will overlap with characteristics of the language abilities we want to measure.
> (Bachman 1990:2)

Thus in language testing 'what is trait and what is method is very hard to distinguish ' (Stevenson 1981:53). The influence of test method *facets* (aspects of the test design) needs to be much better understood, if models of language competence are to be validated.

The problem of empirically verifying communicative competence models is discussed by Hulstijn (1985:373), who asks 'is it [language proficiency] a unitary construct, or does it consist of several subskills, and, if so, what is their number and nature?' He points out that from the psycholinguistic viewpoint,

> since cognitive theories focus on the elementary processes and their integration into routines and strategies, they tend to proliferate the number of hypothesized skills and subskills that play a role in speaking, listening, reading, and writing. Therefore, it is not very likely that such information-processing theories will conceive of language proficiency as a unitary construct, but rather as composed of many skills and subskills.
> (Hulstijn 1985:374)

The issue of the number and nature of language proficiency components has, Hulstijn notes, been raised primarily by language testers seeking to know how many and what kind of tests to administer. He argues that the components in frameworks of higher-order competencies, such as Canale & Swain's, will tend to overlap, being composed of many shared low-order skills.

> The more 'simple skills' are shared by two competencies or 'macroskills', the higher the correlation will be between the scores on the tasks that purport to measure these macroskills, and hence there will be less room to provide empirical support for their independence by means of correlational methods.
> (Hulstijn 1985:377)

It seems that claims that communicative competence 'consists of' three (or four or five) parts, or that it 'includes' linguistic competence may prove to have more pedagogical and practical utility than empirical validity.

For some, communicative competence 'includes' much more: it is a generalized ability to communicate which takes in non-linguistic personality factors: introversion and extroversion, intelligence, experience, etc. This wider understanding of communicative competence accords with Hymes' argument that it is important 'not to separate cognitive from affective and volitive factors' (Hymes 1972), but it can certainly be argued that it is not appropriate for the language tester to attempt to encompass these non-linguistic factors (Alderson 1981b).

Ingram (1985:227) complains of the communicative competence enterprise that 'current research seems more aimed at assessing the nature and construct validity of 'communicative competence' ... rather than devising tests.' As to the role of linguistic competence within communicative competence, things do not seem to have moved on much from the start of the 80s, when Weir noted that their relationship has 'in no sense been clearly established by empirical research' (Weir 1981:30).

2.2 Language difficulty; the development of proficiency

So far we have looked to the fields of second language acquisition studies, teaching and testing, trying to focus the discussion of language ability on models of the mental processes that underly language use. If we now turn to the way language ability develops, we find that the focus changes. This is because a developmental stage is naturally characterised in terms of *what* learners know or can do at successive levels - that is, in terms of the language-related *tasks* that they can perform. The focus thus appears to be more on the *difficulty* of language itself, than on *ability*, understood as a mental mechanism. In

some of the research discussed above the term *proficiency* is used as a synonym for competence, or ability - that is, the mental mechanism; but things would be much clearer if we could think of proficiency as that measure of performance obtained when a certain language ability is confronted with a certain language task difficulty.

We will see how this conceptual model is given expression in Item Response Theory. The remainder of this chapter examines the development of language ability, which includes the notion of language difficulty; it should be clear that, finally, difficulty and ability are mutually-defining notions.

2.2.1 The natural order hypothesis

The discussion so far has suggested a number of constraints or influences on the process of learning - constraints originating in the nature of human cognition, or of language itself, or of the social uses of language. While some of these influences seem to be general, others clearly are particular, depending on the learner (his first language, his learning style), the learning situation (classroom or 'natural'), and the language being learned.

To the extent that general constraints outweigh particular influences, we can expect all language learners to pass through similar stages of development; a 'natural route' will be observable. If particular influences are predominant, we will be more impressed by the variability of interlanguage. It would be most convenient for language proficiency testing if all learners were alike, and their performance in the language to be tested were always completely consistent. This is of course not the case; and yet there still appears to be much evidence in favour of some sort of natural route, or what Pienemann (1985:33) calls

'a universalist perspective', that is, the view that 'all instances of language learning ... might be determined by a set of shared principles.'

That language learning follows its own course - a natural route - in defiance of the best efforts of teachers, is an idea that enjoyed a rise to prominence in recent years, thanks chiefly to Krashen. Along with most of his other ideas it has subsequently come under fierce attack.

Krashen based his hypothesis on two main lines of research: the *morpheme studies* (Dulay & Burt 1973, Bailey et al. 1974, Larsen-Freeman 1975) and error analysis (Dulay & Burt 1972, 1974).

The morpheme studies used a particular elicitation procedure to measure the acquisition of a set of functor words, following similar research methods applied to first-language acquisition. A roughly similar pattern to L1 development was observed. Error analysis seemed to support the view that most errors made by L2 learners could be explained as developmental, being more characteristic of stages of development of the target language than of interference from the first language.

The morpheme studies were criticized on a number of grounds: the findings might be explained as an artifact of the elicitation method (Porter 1977); and the equating of *accuracy of use* with *acquisition order* was questioned (Hakuta 1976). Error analysis has also been criticized. It is frequently difficult to establish the type of error or say why a learner is making it (Schachter & Celce-Murcia 1977); and a simple analysis of errors is inadequate for detecting first language influence (Hakuta & Cancino 1977). Learners might *avoid* attempting structures where because of first language influence they might be likely to make mistakes (Schachter 1974). Additionally, studies continued to demonstrate that first-language influence *does* play a part in second-language learning (Kohn 1986).

The early interlanguage research on which Krashen chiefly drew has thus been largely discredited. But the natural order hypothesis is by no means dead.

2.2.2 Does teaching make a difference?

The natural order hypothesis is at the heart of the debate over whether formal language instruction makes a difference. It is certainly the case that classroom learning is not a simple reflex of classroom teaching, and some researchers go so far as to contemplate

> the possibility that whatever does in fact determine linguistic development in classroom language learners is largely independent of the deliberate teaching acts that are so carefully planned and conscientiously implemented in the classroom (Allwright 1987:210).

Pienemann (1985) points out that if one accepts the hypothesis, one has two logical choices: 'abandon teaching' or 'follow natural order'. Krashen at first recommended the latter option, but in his later writings tends to the former, proposing the *input hypothesis* as a sufficient mechanism to explain how language learning proceeds.

> Humans acquire language in only one way - by understanding messages, or by receiving 'comprehensible input'. ... We move from i, our current level, to i + 1, the next level along the natural order, by understanding input containing i + 1.
> (Krashen 1985:2)

The evidence advanced in support of the input hypothesis has been criticized, and in particular the hypothetical construct of the i + 1 level has been condemned for assuming 'a non-existent theory

of acquisition sequences' (McLaughlin 1987:56). By this he must mean a theory which would would allow one to define a 'level' of language difficulty, and thus to state that a given sample of input contains the level of interest.

Other writers claim that such a theory, if not yet fully worked out, is at least well on the way (Clahsen 1985, for German SLA, Pienemann & Johnston 1987 for English). Pienemann (1985) reviews research into formal and natural L2 acquisition. He finds ambiguous results from a number of studies using *increasing accuracy* as a criterion of measuring progress, or based on morpheme acquisition order, but questions whether these are valid criteria. He discusses a range of other research that suggests similarities between language acquisition in formal and natural settings (Felix 1978, 1982, Hahn 1982, Wode 1981, Pica 1982). He draws a distinction between *developmental* features of language, which follow a sequence fixed by cognitive constraints, and *variable* features which can be influenced by instruction. For developmental features he advances the *teachability hypothesis*, which states that 'at each stage the processing prerequisites for the following stage are developed' (p.37), so that learners can only process (learn) material at the next stage up. His conclusion is thus that a teaching program must follow natural order to be efficient.

Long (1985:86) assents to the idea of acquisitional sequences, but does not agree that these should be explicitly followed in teaching, preferring Krashen's notion of 'rough tuning' language input to the level of learners. He also contests Pienemann's assumption that learners who are at the same acquisitional stage for one structure will be at the same stage in other aspects of their interlanguage development. He cites work on the acquisition of negation (Lamotte et al. 1982) as evidence of 'serious problems for a unidimensional second language continuum with negation as the single predictor.'

Lightbown (1985:105) argues against importing acquisitional sequences directly into teaching, and points out that 'even if every currently described sequence were completely and universally correct, we would still be left with a syllabus sufficient to cover - at most - the first few months of language teaching.'

Pica (1985) compares acquisitional sequences in taught and natural learners, and concludes that classroom teaching can upset natural sequences, accelerating them for linguistically simple features (e.g. the plural -s), retarding the acquisition of more complex features (e.g. the progressive marker -ing), or having no impact on highly complex grammatical items such as the indefinite article.

Sorace (1985) found instruction to have a positive effect on learners with little opportunity to acquire language naturally. Her results indicate that metalinguistic knowledge has 'a more central function than limited monitoring' (p.252), a claim which conflicts with Krashen's.

If teaching *does* make a difference (as Pica and Sorace above both find) the question remains whether it is the *route* or simply the *rate* of acquisition which is affected - that is, is the natural order upset, or merely accelerated? Pica's (1985) finding clearly indicates that the route may be altered. Ellis (1985) also reviews research on the effect of formal instruction. He concludes that of the evidence in favour of teaching making a difference, most concerns the rate rather than the route.

2.2.3 A grammatical view of development

The number of acquisitional sequences for which experimental evidence has been found is small - as Lightbown (1985) points out, far too small to guide the design of a teaching programme. The morpheme order type of study is also unsatisfactory in that

it fails to explain anything about the process of development. It is not reasonable to suggest, for example, that the -ING morpheme is 'acquired' at that point where a learner begins to use it correctly in some obligatory context specified by a particular elicitation device. The -ING morpheme enters into a wide range of grammatical structures, serving a variety of different functions, and a learner's understanding or capacity for use of each of these is not an all-or-nothing affair, but rather a matter of slow transition.

Rutherford says that it is 'difficult to imagine *in what sense morpheme-acquisition research procedures might embrace anything at all in the syntax of the language*' (Rutherford 1987:23, his italics). He criticizes the natural order hypothesis for promoting two (mistaken) assumptions:

> 1. that all of language form is itemizable in the manner of the [investigated] morphemes;
>
> 2. that acquisition of language form is tantamount to steady accumulation of those items in some as yet unidentifiable order.
>
> (Rutherford 1987:23)

Though Rutherford dismisses Krashen's version of the natural order hypothesis as simplistic, his discussion of the development of grammar indicates the operation of general principles that support a universalist viewpoint.

Rutherford observes that grammar serves the creation of *discourse*. Language being a linear phenomenon, discourse demands that information is organized in sequence, in accordance with certain rules (e.g. concerning given and new information). Grammar provides the means whereby information blocks can be ordered within sentences, while remaining interpretable. The effect of more complex grammar is to increase the *distance*

between the syntax and the semantics of the sentence. Thus the notion of increasing semantic-syntactic distance is associated with increasing difficulty, and thus with development.

A learner's early utterances tend to be short, with a simple subject-predicate syntactic structure corresponding to a topic-comment semantic structure. Semantics and syntax are said to be *isomorphic*. In the example sentence

> The war is easy to forget

however, the raising of the object of 'forget' to subject position in the sentence considerably distances the syntax from the semantics (note that the subject contracts no *semantic* relation with the main verb - 'The war is easy...'. Rutherford cites Kellerman (1979) for evidence that

> learners will reject structures such as [the above example] as ungrammatical, *even when such structures are grammatical in the learner's native language*. In other words, the learner will usually prefer that language structure in which syntax and semantics display the greatest isomorphism. To the extent that isomorphism must give way to structure preservation, as is so evident in English, then the challenge posed to the learner is that much greater. (Rutherford 1987:113, italics his.)

Rutherford's discussion of the linguistic devices by which discourse is *grammaticized* seems a promising source of predictions relating grammatical features to the difficulty of test items. However, the semantic difficulty or complexity of discourse is evidently not explicable solely in terms of linguistic features (Brown & Yule 1983).

2.2.4 Linguistic universals

Another area of research which provides some support for the universalist viewpoint, and which also focusses on inherent properties of language, is the work on *linguistic universals*.

Research into language universals has tended to follow either a *data-driven*, bottom-up approach, studying a wide range of languages for evidence of universal patterns and areas of difference, or a *theory-driven*, top-down approach based on the analysis of language to discover the abstract principles of grammar that constrain the form of possible human languages. The first approach is associated with the writings of Joseph H. Greenberg, and the second with those of Noam Chomsky. As McLaughlin (1987:83) notes, 'the Chomskyan approach has tended to assimilate the Greenbergian', and this very brief summary makes little distinction between these orientations.

An interesting example of work on language universals is the *Accessibility Hierarchy* for relativization proposed by Keenan & Comrie (1977). This argues from cross-linguistic evidence that the ease of relativizing noun phrases depends on their sentence function, in the following order (from easiest to hardest): (1) subject, (2) direct object, (3) indirect object, (4) object of a preposition, (5) genitive, (6) object of comparative. The accesibility hierarchy is an example of a chain of *implicational universals*. It states that, for example, if it is possible in a given language to relativize on an indirect object ('the woman to whom he sent the book') it will also be possible to relativize on a direct object ('the child that he hit'). The existence of a feature in a language implies the existence of all features higher in the hierarchy.

In the usage of typological studies, element (1) in the hierarchy is the least *marked*, and element (6) is the most marked. Markedness is a key concept in the study of universals. Here it is seen to be a matter of degree, the most unmarked form being the 'most natural' or 'most universal'.

Examples have been found of languages for which this prediction is not true, and so it appears preferable to give it the status of a *tendency* or *statistical universal* rather than an *absolute universal* (Comrie 1984).

Comrie (1984) refers to a 'vast amount' of work within second-language acquisition studies that has shown how the conclusions of the Accessibility Hierarchy 'translate fairly directly into valid predictions about the acquisition of relative clauses in a second language.' But he notes areas of less than perfect fit which indicate that other factors must also be admitted:

> The claims of a literal psychological interpretation of the Accessibility Hierarchy hold only where other things are equal and does not exclude the possibility that other factors, for instance processing strategies or real world likelihood of interpretations, might at times override the predictions made by the Accessibility Hierarchy on its own. (Comrie 1984, 19)

This is of course reasonable, and suggests that efforts to explain similarities or differences in second-language development using theories of linguistic universals will be at best only partly successful. Applications to second-language acquisition of the notions of *markedness* or *parameter setting* in Universal Grammar theory are, at best, premature. Kean (1986) warns against the straightforward importation of specifically linguistic concepts into second-language acquisition research. In a discussion of the distinction between the *core* and *periphery* in Universal Grammar, (which corresponds essentially to the

distinction between unmarked and marked features of grammar) she argues that given the dynamic nature of interlanguage grammar, comparisons of markedness in L1 and L2 are inadequate to predict performance. Transfer is certainly not only influenced by markedness, says Kean, and ignoring other factors yields data which is 'wildly inconsistent and intractable in terms of rational theory construction' (Kean 1986:90).

2.2.5 Variability and first language transfer

Interlanguage variability has many sources, and not all types of variability undermine the universalist viewpoint. The following discussion is based on a categorization by Ellis (1985):

1. Systematic variability
    1. Individual variability
    2. Contextual variability
        1. Linguistic context
        2. Situational context

2. Non-systematic variability
    1. Free variability
    2. Performance variability

(After Ellis 1985:76)

Performance variability is non-systematic variability due to random errors, and is not significant of underlying interlanguage states or processes.

Free variability is the apparently random use of two or more language forms in one or more particular linguistic or situational contexts. Ellis sees it as essential to the process of hypothesis testing which underlies learning; in time, forms

become firmly mapped onto particular functions. Free variability may make learners' performance harder to interpret, but is not inconsistent with a universalist position.

Systematic variability can be observed not only in formal features of interlanguage, but in the way it is used as well.

Individual variability is that systematic variability which occurs to the extent that learners' interlanguage is not constrained by internal factors (cognitive processes, universal grammar) or external factors (the nature of the L2, their experience of instruction). Learning styles vary, as do the strategies that learners apply to organizing and restructuring their knowledge. This *does* undermine the universalist position, of course, although Ellis (1985) says that most evidence of individual differences concerns *rate* of learning rather than the *route*.

Contextual variability is systematic variability which is dependent on linguistic or situational context. Ellis (1985) provides evidence of learners systematically supplying a target-language form in one linguistic context but not in another. This kind of variability exposes the shortcomings of the morpheme studies, but is entirely consistent with the universalist position. In a language test, manipulating context allows the difficulty of an item to be adjusted.

Variability across situational contexts is the phenomenon investigated by heterogeneous-competence (or *multiple-competence*) models such as those of Ellis (1987) and Tarone (1983). In these models it is differing degrees of *attention to speech* that produce styles from the *vernacular* to the *formal*. Paradoxically, it is the formal style which is *least* stable, presumably because it represents more recently acquired, less automatized knowledge, and conscious appeal to the first language, or other

communication strategies. Tarone (1987) discusses the factors which affect attention to speech: the interlocutor, the topic, the task, the amount of discourse generated, time pressure etc.

An important source of variability, and one which certainly complicates language testing, is the influence of a learner's first language.

Behaviourist learning theories called this influence *transfer*. Some of the original statements of Interlanguage theory retained first language transfer as one of the important influences on interlanguage development (Selinker 1972). Subsequently, learners' errors were argued to be largely developmental in origin, and the significance of transfer was minimized (Dulay & Burt 1972).

Presently the phenomenon of transfer is again recognised, and research into language typology and linguistic universals has kindled fresh hopes of being able to explain it. Notions of markedness and parameter-setting may prove more adequate as predictors of transfer than the traditional contrastive analyses which were largely discounted during the heyday of error analysis. Kellerman & Sharwood Smith (1986:3), who propose the term *cross-linguistic influence* (CLI) to replace that of transfer, with its behaviourist overtones, call CLI a 'pervasive phenomenon in second-language acquisition'.

In testing, item bias relating to L1 influence has been detected (Chen & Henning 1985, Pollitt & Taylor forthcoming). The present study will also look at this question (5.3.3.5 below).

2.2.6 The teaching viewpoint: organisation and grading

If we are interested in how language ability develops within a formal instructional setting (and the present study is), then we cannot neglect the teaching viewpoint. After all, decisions

concerning syllabus design, grading and sequencing all reflect a pedagogic view on what is easier or more difficult, or on what is appropriate to different stages of language development. Of course, there never has been a single orthodoxy in language teaching, and nowadays, to the extent that the pursuit of communicative competence has been widely espoused as the proper goal of language teaching, the variety of approaches is wider than ever.

Canale & Swain (1980) review existing 'communicative' approaches to teaching and testing. One general approach they identify emphasizes attaining a minimal or 'threshold' level of communication skills, in order to survive in a range of common second-language situations (e.g. Van Ek 1976). They are generally critical of this approach, finding the notion of a 'minimum competence' ill-defined, and rejecting the stress on communication, in the sense of getting one's meaning across, at the expense of grammatical accuracy. They note that such approaches neglect situational appropriacy.

Canale & Swain also consider more theoretical, sociolinguistic approaches (e.g. Hymes 1967, Halliday 1973, Allen & Widdowson 1975, Munby 1978). Syllabuses based on these approaches are criticized for an 'overemphasis on communicative functions' (p.23) as an organizing principle, with the factors of grammatical complexity and transparency being neglected. Other writers have criticized this tendency for communicative syllabus design to be realized through the wholesale importing of categories from sociolinguistics, with a consequent increase in the *content* of teaching programmes:

> When we speak of 'communicative language teaching' we are (in common usage) referring to one which ... bases itself on inventories specifying conceptual and pragmatic categories which are arrived at by considering presumed communicative needs.
> (Johnson 1982:122)

Johnson calls this the 'teaching content' solution, and argues that it necessarily reinforces the view of language proficiency as a collection of discrete bits of knowledge (what Rutherford (1987) calls the 'accumulated entities' view). Johnson points out that there is an alternative interpretation of 'communicative': one which refers less to syllabus content, and more to methodology. Prabhu (1987) proposes the terms *communicative* and *communicational* to distinguish between these two senses.

A 'communicative' syllabus, then, is most often one in which semantic and sociolinguistic categories - notions, functions, topics, situations of use etc - are introduced alongside the more traditional grammatical and lexical inventories. The overriding practical problem, as Canale & Swain (1980) noted, is to find an organizing principle to guide the selection, combination and sequencing of items from these various categories.

In a structurally-organized syllabus the organizing principle is the language system. First a grammar point is identified, and the other elements - notions and functions - are selected to provide meaningful contexts for presenting and practising that grammar. With a traditional structural syllabus it appears to be fairly straightforward to select grammar points in order to achieve a progression from simple to complex. The same cannot be said for syllabuses where the organizing principle is notions or functions.

> A strict linear shape does not work well when the categories of language content are notional or functional since there is no inherent sequence or order in them which seems best. (Dublin & Olshtain 1986:51).

> There is one central and persuasive argument against the use of functional syllabuses at the zero beginner level. It is simply that a functional organization automatically implies structural disorganization.

(Johnson 1982:107)

Of course, the ordering of elements in a structural syllabus is never done simply on the basis of linguistic simplicity, however conceived. Other traditional principles - frequency of occurrence, valency - reflect the assumption that the most learnable progression is one which takes into account the utility of grammar, not merely its structure. Thus traditional structural syllabus design is, necessarily, illuminated by pedagogic insight and intuition.

None the less, there is an assumption that the organization of a structural syllabus is in some real sense *intrinsic* - that some 'inherent sequence or order' can be found in grammar itself. According to Brumfit (1981) this makes it inherently superior to syllabuses organized around notions or functions. In a published exchange, Brumfit and Paulston both take issue with Wilkins, one of the original proponents of notional-functional syllabuses, arguing that such syllabuses lack intrinsic organization and hence cannot be related to any theory of language acquisition. Wilkins argues in return that the organization of structural syllabuses is no less extrinsic:

> The only intrinsic ordering [for a grammatical syllabus] that I could conceive would be one that had psycholinguistic validity. This is an area where it has proved notoriously difficult to cast any light on the relative status of grammatical categories or rules.
> (Wilkins 1981:99)

As we have seen in the previous section, there are writers who claim to have made progress in this area. Clahsen (1985) and Pienemann (1985) both make concrete proposals for grading based on acquisitional sequences. Most writers however stress that too little is known to make such recommendations. In their introduction to the collection which contains the papers just mentioned (Hyltenstam & Pienemann 1985) the editors declare their

intention to 'counteract the emergence of a 'psycholinguistic' or 'developmental' method for foreign/second language teaching.' In the same collection Lightbown stresses:

> We are still at too early a stage in our understanding of how natural acquisition sequences can or should be related to teaching sequences to make specific recommendations for 'grading' or sequencing.
> (Lightbown 1985:103)

None the less, it *is* true that grammar lends itself to hierarchical classification much better than do notions or functions (Dublin & Olshtain 1986). It is also undeniable that a sort of pedogogic consensus has grown up in the English language teaching field, so that influential coursebooks, reference materials and public exams tend to cover a great deal of the same ground in roughly the same order and using the same conventional structural units - ('the conditionals', 'modals', 'some and any' etc). If formal instruction has any influence at all, then it is likely, at least, that learners will not know what they have not been taught; and thus pedagogical notions of difficulty *may* indeed be reflected in any general measure of language proficiency.

2.2.7 Language development and cognitive difficulty

A *procedural*, or *task-based*, syllabus is one which is organized in terms of learning tasks graded for cognitive difficulty. The language necessary to perform the tasks is not specified, the assumption being that it adjusts itself automatically to the demands of the task. This is the recommendation of McLaughlin, discussing the relevance of second-language acquisition studies to grading grammatical materials. He cites Corder:

> The progressive elaboration of the interlanguage system of the learner is a response to his developing need to handle even more complex communicative tasks. If we can control the level of these correctly, the grammar will look after itself.
> (Corder 1981:78, quoted in McLaughlin 1987:164)

Long (1985) makes a similar recommendation, seemingly associating *tasks* with professional needs. Prabhu (1987) describes extensive applications of procedural syllabuses. Prabhu's work is particularly interesting. Apparently carried on with little awareness of Krashen's concepts of *acquisition* and *comprehensible input*, it nevertheless bears strong similarities. What is striking, and what indicates problems with exploiting this approach to difficulty in testing, is that the notions of simplicity or comprehensibility cannot be related simply to the language involved in performing a task, because they are rooted in the whole context, which includes the non-linguistic means used to negotiate meaning. As Long (1985) notes, comprehensible input is made comprehensible by adjustment not merely of language, but of the whole context of speech. Furthermore, *comprehension* is a matter of degree: it can be defined only relative to the task at hand (Prabhu 1987).

Mention might also be made here of a system proposed to describe all educational achievement in terms of a common process of cognitive development. This is the SOLO (Structure of the Observed Learning Outcome) taxonomy (Biggs & Collis 1982). Biggs & Collis, starting from Piaget, identify five developmental stages: pre-structural, uni-dimensional, multi-dimensional, relational and extended abstract. They adapt Piaget's notion of stages, in that they view them not as innate properties of the person, but as reflections of educational attainment, amenable to teaching and possibly differentially developed in different school subjects. They then set about characterizing the kinds of

behaviours which indicate, for a particular school subject, the developmental stage reached. Modern languages are included in the examples they give. A couple of examples will be useful.

Given the task of translation from L2 to L1, they characterize a word-for-word approach as unistructural, a translation in which the odd functional word is changed as multistructural, and a good, free translation as relational.

The following is an example of a task where the student must 'find the rule': a list of French sentences with *est* or *soit* is given, and the task is to state the rule for the subjunctive, and complete two blanked sentences (pp. 151-152).

A relational response: '*Est* is used when you are positive about something and *soit* is used when you are not sure about something.' This is described as failing to seek a general hypothesis against which the student can test all sentences. A better, extended abstract response might be:

'*Soit* is used after impersonal expressions that indicate the personal opinion or doubt of the speaker.' This response shows that the student has 'gone beyond the immediate content to set up hypotheses and has used the data to test them.'

Discussing methods of teaching, Biggs & Collis characterize the *audio-lingual* approach, with its emphasis on habit formation and learning by analogy rather than analysis, as being geared to the unistructural/multistructural level of response where the student is not specifically encouraged to see a relating principle in the stimuli presented. Thus, they reason, it is satisfactory for beginning but not for higher levels of language functioning.

The *audio-visual* approach adds aids to comprehension, and thus encourages development of relational responses.

But 'the very top level of functioning' in SOLO terms is the extended abstract. In languages, this would be demonstrated by ability to perform the kind of philological grammatical analysis illustrated above. They suggest that secondary schools should aim at a lower, (i.e. relational) level, thus releasing children and teachers from 'unrealistic curriculum aims.'

Professionals in the language teaching or testing community might query the relevance of the SOLO taxonomy for describing proficiency. The kind of scholarly ability described above seems to be of particularly marginal importance for most learners.

Cummins (1980) makes a different link between language proficiency and cognitive difficulty. He claims that the *developmental* dimension is missing from the Canale & Swain model of communicative competence. Humans start off, he states, with a 'species minimum' of linguistic competence. Communicative competence develops from experience of living in society. Most people, given suitable exposure, will develop what he calls BICS - basic interpersonal communicative skills. A higher competence is CALP - cognitive-academic language proficiency. This is an *analytic* competence: the prolonged operation of thought processes on linguistic representations. This competence is encouraged by formal education, and is, indeed, necessary in order to achieve academic success. Language is here seen as a *tool* of scholarship, not its *object*, as in Biggs & Collis' discussion of modern language teaching.

Cummins (1983) modifies the CALP/BICS dichotomy, identifying two factors: the *range of contextual support* and the *degree of cognitive involvement*. Uses of language where there is much non-linguistic context to aid communication, as in most face-to-face interactions, are called *context-embedded*. Uses of language where there is less contextual support, like writing an essay or reading an academic text, are called *context-reduced*. Context-reduced language use is thus generally speaking more difficult. Communication is *cognitively demanding* to the extent

that it requires active concentration. The degree of cognitive involvement required for a particular task thus depends on the speaker's level of language proficiency - the lower the level, the more concentration is necessary - and on the nature of the task: generally speaking, more academic tasks are more difficult.

2.3 Discussion

2.3.1 General language proficiency

This chapter began with the claim that in order to construct an objective measure of language ability it is necessary to impose the qualities of unidimensionality and invariance upon it. Spolsky's notion of General Language Proficiency (GLP) appears to do just this; and yet given the complexity of current models of communicative language ability, we must ask how much is gained or lost by opting for this simple notion.

Davies distinguishes two basic approaches to General Language Profiency:

> First, there is the philosophical argument: this may be what is meant by construct validity if it allows testing. ...Second there is the competence-performance argument. Since this is either a philosophical or a practical issue (ie we are testing one or the other) this merges into one of the other arguments. Third, there is the practical argument ... which says in view of our lack of clarity it is best to gather as much evidence as possible from a wide variety of tests. (Davies 1981:185)

Oller's expectancy grammar is the prime example of the first approach: GLP as a unitary competence. Davies represents the 'practical argument'. But we should avoid the conclusion that the central issue in discussing GLP is the psychological reality

(or otherwise) of some central language competence. The 'realism' (Cronbach 1988) or otherwise of GLP or of any such construct should not be an issue.

> The question of whether or not we believe these constructs actually exist, in some physical sense, in our brains, is not relevant to construct validation.
> (Bachman 1990:292)

The discussion is thus, finally, about different measures: is there some single measure which can be held adequately to characterize a learner's general proficiency? Any answer to this question must state what is meant by *adequately*; that is, it must recognize that assessments are made for a particular *purpose*, and the crucial issue is the range of valid inferences which can be made from any GLP measure.

Davies is sceptical about the possibility of validating the GLP construct, because of 'our lack of clarity' about the underlying processes of language. Similarly, Ingram (1978) discussing the 'disjunctive fallacy' (that discrete-point and integrative tests cannot both be valid), argues that both are needed, given our imperfect understanding of how either of them works. Following this argument, General Language Proficiency is to be understood as meaning *overall* language proficiency; that is, a picture built up by aggregating *different* measures of proficiency.

Several writers define proficiency as the ability to use language to some purpose. For example, proficiency is 'not just knowlege but the ability to mobilize that knowledge in carrying out particular communication tasks in particular contexts or situations' (Ingram 1985:220). Or it is 'how successful the candidate is likely to be as a user of the language in some general sense' (Morrow 1981:18). As Hughes (1981:176) points out, to test proficiency we are at liberty to choose whatever language-based tasks we like - 'solving anagrams, finding rhymes, judging the grammaticality or acceptability of sentences, making

translations, or even doing cloze tests.' But certain tasks seem more central to our interests. Choosing relevant tasks, we *choose* our definition of proficiency, and this choice reflects our view of the value of language teaching in general, or the purpose of a particular course. If we consider a practical ability to communicate in a foreign language to be the most desirable outcome of studying it, then we may well choose communicative tasks as our measure of proficiency. One might claim that a communicative test is a *better* test because one believes that communication is a better goal of study, but it is not clear that measuring communicative ability gives a better or in some way truer picture of the mental apparatus that constitutes a learner's competence.

The practical problem with selecting tasks for assessing proficiency is to decide which tasks are sufficiently representative of the (doubtless complex) specification of proficiency we have chosen. If one task were completely representative of all aspects of proficiency we considered important, then we would only need one test. In practice this is rarely the case. Proficiency on a listening test will probably differ from proficiency on a writing test. If we consider that both listening and writing are important, we need both tests, or we lose information and do someone an injustice.

Each test measures a specific (and let us assume relevant) proficiency. Aggregating or averaging the results of a large number of different tests obviously gives the best possible measure of GLP, in the sense of overall proficiency being discussed here.

We saw that cloze and dictation found popularity because they seem to approximate better to this best possible measure than do other tests. Certain tests, notably of speaking in the traditional four-skills approach, tend to approximate more poorly than others, and might thus be considered worse candidates as measures of GLP.

In this view, to claim that a certain test (cloze, say) is a good test of GLP is simply to imply that it gives an acceptable approximation to the result you would get if you administered a large number of different, relevant tests. It need not be a claim that the test directly illuminates the nature of language competence.

It is sometimes argued that the nature of competence is of no interest to testers anyway. What is of interest is a learner's ability to *do* certain desirable things; therefore a good test should simply require the candidate to do those things. Ebel (1979) argues:

> Most of what we teach in educational institutions are knowledges, skills, and abilities. These can all be defined operationally. They are not hypothetical constructs.... We would speak more sensibly, I think, if we did not call them constructs.
> (Ebel 1979:307)

Or again:

> There is no better way of making clear what one means by achievement in algebra or chemistry or psychology than by describing how one would measure the amounts of those achievements that other persons possess. ... Good tests of human traits can provide useful operational definitions of those traits.
> (Ebel 1979:301)

Ebel would consider that GLP can be given a satisfactory operational definition by the choice of relevant tests. He would not consider it likely that tests could throw light on the psychological processes underlying language performance, nor even necessary that they should attempt to.

And yet we cannot get away from the need for construct validation: for theories about language competence. This is not simply because it seems rather unsatisfactory to use, say, cloze tests, when 'it is a fact that no one has a clear idea of just what a cloze test is measuring' (Farhady 1983:256).

The notion of a 'relevant test' used above sounds simple but conceals a problem: how does one decide what is relevant and what is not? Some proficiency tests are evidently relevant because they directly test performance on some primary goal of learning - a test of spontaneous speaking skills, for example, is relevant to assessing a communicatively- oriented course. But if we decide that a grammar test - for example - is relevant, then we do so (if not out of mere attachment to tradition) from a conviction that grammatical knowledge is an important element in achieving final goals. Grammar is one of many possible *enabling skills*. This conviction must be rooted in some theory of how language is learned, or how language competence is structured.

The reason for including enabling skills in the range of relevant proficiency tests, rather than confining one's attentions to performance tests of final goal behaviours is that, being more central, they provide more *generalizable* measures. Weir, questioning the notion of performance tests, reasons:

> A performance test is a test which samples behaviours in a single setting with no intention of generalising....Any other type of test is bound to concern itself with competence for the very act of generalising beyond the setting actually tested implies some statement about abilities to use and/or knowledge.
> (Weir 1981:30)

In other words, there are general competences that underlie particular performance in particular situations. Ingram (1985) reviews the problems involved in relating observed behaviour to more generalized competences.

First, the *redundancy* inherent in language, and the fact that skills can develop at different rates, mean that different learners may tackle a particular task in different ways, calling on different aspects of competence. It is probably impractical to specify *which* enabling skills are necessary to the performance of a particular task, or the *relative importance* of such skills as are necessary. (Alderson 1981:49).

Secondly, it is by no means clear what general *range* of tasks are tested by testing performance in a particular task. 'Somebody who has never used public transport, for instance, may ... fail to carry out the necessary tasks readily or appropriately if he has to buy a bus or train ticket ... even though he has mastered such 'functions' as seeking information.' (Ingram 1985:222).

Thirdly, there is the problem of generalizing from one *situation* of use, or *topic*, to another. This is particularly an issue in ESP (English for specific purposes) testing. Is any academic text equally suitable for measuring the reading proficiency of all academics? Or should historians be given history texts? And if so, on what period of history? The ELTS exam offered modules for six subject specialisms, but on very questionable theoretical bases (Criper 1981).

Thus the more specific and authentic a testing task is, the greater the problem of generalizing to other tasks, and thus to overall proficiency. Hence the attraction of less specific, less authentic tasks which come closer to the 'heart' of language competence. Davies is in no doubt about what the deepest-lying, most central competence is:

> What remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that grammar is at the core of language learning... Grammar is far more powerful in terms of generalisability than any other language feature.

(Davies 1978)

This is, we might say, the traditional view. Carroll (1983:94) seems to be making a weaker claim than this, when he explains the existence of the General Language Proficiency Factor like this: 'a language is a language.... That is, a language is an interrelated system.' Carroll reasons that all non-trivial use of language simultaneously exercises a variety of different competences, causing them to develop in harmony. The General Factor is the *result* of the way language is used; it is not evidence of a *causative* unitary competence (as proposed by Oller). But in the same article Carroll shows that he too believes language competence to have a centre. He cites his own (1966) discussion of one particular analysis of a test battery:

> It is not surprising that the four skills tests should be found to measure primarily a singlngle factor of language proficiency in common. Basic competence in a language - knowledge of its phonology, morphology, syntax, and lexicon - is required by each of the tests, no matter what particular 'skill' it measures. The high loading of the writing test on the common factor may reflect the fact that this test is probably most demanding with respect to the morphology and syntax of the language. Many of the other tests appear to demand knowledge primarily of lexicon, which some would regard as less close to the heart of language structure.
>
> The fact that the speaking test is least associated with the common factor of overall language proficiency may indicate that the requirements of the task set by this test are fairly specific and possibly to some extent unrelated to the measurement of language proficiency...
>
> (Carroll 1983:95)

Here, the common factor is held to reflect the workings of a common core of language competence associated with morphology and syntax, and we seem to be given *two* explanations of why tests

should load less heavily on this common factor: the tests that demand knowledge of lexicon do *not call on* the core competence (they are 'less close to the heart of language structure'); the speaking test, on the other hand, calls on specific skills, presumably *in addition* to the core competence.

If grammar is the centre of language competence, then does this include explicit knowledge of grammar? Carroll seems to say as much when he describes language learning as 'a process of acquiring conscious control of the phonological, grammatical and lexical patterns of a second language, largely through study and analysis of these patterns as a body of knowledge' (Carroll 1966: 102). But the utility of studying grammar, except to boost proficiency in grammar tests, remains a contentious issue, precisely because there is still no way of telling how explicit knowledge relates to implicit linguistic competence, nor yet how linguistic competence relates to communicative competence (Weir 1981).

Latterly, given the interest in communicative measures of proficiency, there is a tendency to discount the relevance of tests which appeal to explicit knowledge of grammar. Ingram (1985) is typical:

> Knowledge and proficiency are not the same: one can have much knowledge about a language and even be able to recall and consciously apply many grammatical rules and yet not be proficient in the sense of being able to utilize that knowledge readily for practical communication purposes. (Ingram 1985:219)

This statement demonstrates clearly enough that what constitutes proficiency is a matter of declaration (Ingram is saying that grammatical knowledge is not proficiency because proficiency is something else). Ingram goes on to report research which demonstrates that 'the level of correlation between tests of formal knowledge and tests of practical proficiency seems to

depend on the nature of the course or the environment in which the language has been learned.' In China, for example, where English is learnt largely through formal study, with little opportunity for natural acquisition, tests of formal knowledge correlate very poorly with tests of practical proficiency. But this should not justify the conclusion that the knowledge of English possessed by Ingram's Chinese subjects is marginal - is not 'real' competence. Compared with a fluent speaker of the language they are certainly *missing* something, and they may well know something useless that the fluent speaker doesn't; but I do not see how this kind of study can show what linguistic competence they and the fluent speaker have *in common*. Again we see the difficulty of inferring the structure of competence from the structure of proficiency.

But Ingram, and others who follow Krashen in distinguishing learning and acquisition, are of course starting off with a particular model of the structure of competence: in this view *acquired* knowledge is the real core of language competence, and *learned* knowledge is something marginal or external, whose influence in testing must be minimized if we are to have a true picture of a learner's proficiency. Thus Ingram (1985:234) recommends that tests should be so framed as to minimize monitoring, 'except perhaps in writing'. This qualification implies, generously, that learners may be given more than sixty seconds to produce an essay. Otherwise, 'the learning-acquisition distinction and the concept of monitoring mean that the interpretation of results on indirect tests must take into account the time allowed.' This insistence that learning (in Krashen's sense) is not learning (in the popular sense) follows logically from the following chain of reasoning: Acquired knowledge is the heart of language competence; acquisition proceeds according to developmental sequences; the proper measure of language proficiency is in terms of the developmental stage reached; any test measuring something other than acquired knowledge will give a biassed picture of proficiency - that is, a wrong impression of the learner's developmental stage.

The test which Ingram has worked on is the ASLPR (Australian Second Language Proficiency Ratings), which is a development of the United States Foreign Service Institute School of Language Studies (FSI) scale. Ingram claims that the ASLPR rating scales are based on developmental criteria. The scales produce a four-skills profile. Ingram reports frequent cases of quite widely separated profiles, using this as evidence against the Unitary Competence Hypothesis; development, we are to understand then, can proceed at different rates for different skills.

The Interagency Language Roundtable (ILR) oral interview (Lowe 1982) and the *ACTFL Proficiency Guidelines* (American Council on the Teaching of Foreign Languages 1986) with the oral interview test based on them, adopt a similar approach. Lowe (1988) defines proficiency as follows:

> proficiency equals achievement (ILR functions, content, accuracy) plus functional evidence of internalized strategies for creativity expressed in a single global rating of general language ability expressed over a wide range of functions and topics at any given ILR level. (Lowe 1988:12)

This developmental basis for proficiency testing is certainly appealing, but does it provide adequate criteria for test construction? J.H. Hulstijn (1985) criticizes Ingram's unqualified preference for direct tests like the ASLPR, pointing out that 'any test is a "trait-method unit",' that is, that a direct test is a test like any other, and introduces method effects. He cites Alderson: 'we must give testees a fair chance by giving them a *variety* of language tests' (1981b:190); Alderson here is summarizing the view of many.

'Proficiency is what proficiency tests measure' - that is, proficiency is an operationally-defined construct, a measure arising from performance on test tasks. Test tasks are framed to

be relevant to the chosen purposes of study or a view of how language is learned. General language proficiency can be understood as an average from the whole range of relevant proficiency tests. To narrow this range to manageable limits it is desirable to select tests which produce generalizable results. This can be understood in two ways: integrative tests such as cloze provide good measures of GLP because they appeal simultaneously to a wide range of skills. Other tests, for instance, of grammar, may be preferred because they are believed to measure a more general, central aspect of language competence. One example of the first, global view is the use of rating scales to operationally define proficiency, as in the ILR or ACTFL interviews. Lowe (1988:14) contrasts this 'holistic, top-down view' with the 'atomistic, bottom-up' view underlying communicative competence models. Thus assumptions about the nature of language competence are important to the selection of relevant proficiency measures. Presently there are conflicting views of the nature of language competence, hence conflicting recommendations as to relevant proficiency measures, even if there is general agreement that the most desirable outcome of language study should include communicative competence. The models of communicative competence so far offered do not appear to clarify the place of linguistic competence within the whole; but they provide a useful framework for defining language proficiency in terms of both language tasks and hypothesized language competences.

2.3.2 Conclusion

Thus we can identify two contrasting views of core competence: a linguistic view, placing grammatical knowledge at the centre, and a developmental view which minimizes the relevance of grammatical knowledge as far as this is explicit and at odds with the learner's current developmental stage. Ingram, whose advocacy of 'direct' tests has been mentioned above, concedes that indirect tests might also be structured to establish what point in the

developmental scale the learner has reached, although for this to be possible 'a common developmental schedule has to be identified and sufficiently clearly delineated to make test construction feasible' (Ingram 1985:234).

The present study sets out to construct a proficiency trait using discrete items that touch on traditional pedagogic problems. This follows directly from the item bank's intended role as an instrument for *formative* assessment within a teaching programme. Because many items have a clear pedagogic point, performance on a test can provide a detailed recipe for remedial action on the part of individual learners, and the item bank as a whole can provide a detailed picture of what it is that learners typically know at different levels. Item bank tests are unspeeded, lack 'authentic' communicative purpose, and allow the testee recourse to explicit, conscious knowledge. The present study shows that a reasonably coherent language proficiency trait can be constructed in this way, but in the light of the above discussion, there are evident dangers in interpreting such a trait in 'developmental' terms.

There are also clear limits to the range of valid inference that can be drawn from performance on such indirect, competence-oriented tests. As a *summative*, or end-point, test, in any teaching programme where a practical ability to communicate is an important goal, the competence-oriented test is inadequate on its own, and must at least be complemented by relevant skills-oriented tests. Formative testing is different. In an instructional setting future outcomes are of greater concern than present payoffs, and it can be argued that the competence-oriented test may be more revealing of learners' language development. Higgs & Clifford (1982) looked at the consequences of different proficiency profiles for learners taking the FSI/ILR interview procedure, they found that learners with even profiles (on scales of grammar, vocabulary, fluency etc) were more likely to continue making progress than learners with higher ratings on vocabulary and fluency relative to

grammar. The latter pattern was associated with the onset of fossilization. This throws an interesting light on the significance of grammatical competence, and perhaps allows us partially to reconcile such contrasting positions as those of Davies and Ingram. While grammatical knowledge may not relate strongly to present performance, it may well be a good predictor of future performance, as long as other conditions are satisfied.

This chapter opened with the statement that in order to measure a psychological trait such as language proficiency, it is necessary to *impose* the quality of unidimensionality on it. It should be clear, then, that the present attempt to construct a proficiency trait through item banking does not in itself presuppose a strong claim about the underlying nature of language competence.

> Unidimensionality ... is a psychometric property independent of any concept of 'dimensions' of language proficiency, which are psycholinguistic properties or concepts (Hamp-Lyons, 1989:115).

Hamp-Lyons makes this useful point in the course of a discussion which otherwise appears to demonstrate some misunderstanding of this very issue (see below, 3.3.2). As McNamara (1990:112) stresses, it is essential

> to distinguish consistently between two types of model: a measurement model and a model of the various skills and abilities potentially underlying test performance. These are not the same thing.

We can refer this back to the discussion above. The construct of general language proficiency may amount to a strong claim about the 'skills and abilities underlying test performance', but may be based on a weaker, 'measurement model' claim.

The basis for selecting items for the present bank is the latter, weaker view of general language proficiency. It is as inclusive as possible; the items are quite heterogeneous in terms of content. Taking Bachman's (1990:87) taxonomy of components of language competence, we might say that the following are all addressed:

> all components of grammatical competence except phonology (vocabulary, morphology, syntax, graphology);
>
> all components of textual competence (cohesion, rhetorical organisation);
>
> certain compenents of illocutionary competence (ideational functions, manipulative functions);
>
> certain features of sociolinguistic competence (sensitivity to register, to naturalness).

The fact is of course that most of these 'components' could not in any case be satisfactorily separated from each other for testing, given that 'in language use these components all interact with each other and with features of the language use situation' (Bachman 1990:86). At the same time the nature of discrete-item paper-and-pencil testing means that the item bank is able to address certain components more squarely, let us say, than others: vocabulary, morphology and syntax, or the components of grammatical competence (in Bachman's scheme).

The hypothesis is that a fairly heterogeneous collection of items, many relating to traditional pedagogical language problems, *can* be fitted satisfactorily to a unidimensional trait: that is, that a language proficiency trait defined in these terms can be measured. To say that it *can* be measured is not yet to say that it is worth measuring; that is the question addressed by the investigation of item difficulty in Chapter 6. To the extent that item difficulty is explicable in terms of

language and language use (linguistic, psycholinguistic or sociolinguistic factors), rather than in terms of factors internal to the test, then the trait becomes interpretable, and we are better able to judge what valid inferences may be drawn from performance on item bank tests.

The selection of items for the bank is discussed at greater length in Chapter 5. The thorny issue of unidimensionality (one might almost say *hoary*, borrowing Hamp-Lyons' (1989:114) epithet for the unitary/divisible competence issue) will be taken up in the following chapter, where Item Response Theory is presented.

In the concluding discussion (Chapter 7) we shall consider to what extent the language proficiency trait depicted by the bank can be interpreted in developmental terms. Let us end this chapter with Swan's (1987:66) warning against understanding 'development' too narrowly: attempting, that is, to derive

> a very general view of language use and development from limited data of a very particular kind - from those phonological and grammatical features which do exhibit variability. This sort of perspective might lead one, for instance, to say that a learner's interlanguage had entered a new stage of development because of alterations in the pattern of variability of a few phonemes and morphemes, but to treat his/her acquisition of 2,000 new words as developmentally unimportant. ... Variability research is typically concerned with that rather special category of problematic linguistic elements which tend not to be mastered successfully: those phonological and syntactic features which learners find especially difficult, and where competing interlanguage rules or habits lead to variability. While research can obviously contribute to our understanding of such matters, it is important not to overestimate their importance in language development. (Swan 1987:66)

3: Item Response Theory

3.1 The construct of language proficiency

The previous chapter examined the notion of language proficiency, attempting to draw a distinction between mental models of language competence, and the language difficulty which resides in situations of language use, or tasks. As the parameters of *ability* and *difficulty*, these strands find explicit expression in Item Response Theory (IRT). Together they can be taken to define the *trait* of language proficiency. Vollmer (1981) in his discussion of the concept of General Language Proficiency, mentions

> a criticism developed by the sociological school of the Symbolic Interactionism against the traditional trait concept and picked up by interactional psychology,... that the unit of analysis in the behavioural sciences cannot be the structure of human capabilities (the assumed stable 'traits') but will have to be the interrelationship between task situation and persons involved. (Vollmer 1981:164)

Language proficiency is a trait in this sense. We think of it as an attribute of people, but it can be defined only in terms of tasks in a test of some kind. Vollmer's (1981) much quoted dictum that 'proficiency is what proficiency tests measure' might be intended as a criticism, but it is in fact a simple definition: competence is a psychological construct, but proficiency is a *measure*.

This chapter introduces Item Response Theory, and shows how by selecting tasks (test items) we may attempt to construct a unidimensional ability-difficulty trait, extending from a low to a high level of proficiency. When the trait is constructed,

there remains the need to demonstrate that it has coherence, and measures what we would like it to measure. This is the object of *construct validation*. This chapter discusses how construct validation can be pursued in IRT.

3.2 An introduction to Item Response Theory

3.2.1 Shortcomings of standard testing methods

It is customary to begin an introduction to IRT by discussing the shortcomings of traditional approaches to testing which IRT claims to address (Wright & Stone 1979, Henning 1984, Hambleton & Swaminathan 1985).

There is the problem in classical test theory (CTT) of the generalizability of test scores beyond the sample of persons tested. The significance of item difficulty and discrimination is dependent on the persons sampled, and conversely, the significance of person ability measures is dependent on the items.

Reliability and validity estimates are likewise sample-dependent. Referring to the test as a whole, they are accurate only for scores near the mean, whereas error of measurement is greater as scores depart from the mean.

CTT's fundamental concept - reliability - depends on *parallel forms*, which are in practice difficult to achieve.

Most importantly, CTT cannot deal with test difficulty by comparing individual person ability with particular item difficulty, but rather looks at group means and distributions. It cannot 'quantify the appropriateness of a given test for a specified individual, nor... the appropriateness of a particular item for inclusion in a particular group of test items' (Henning 1984:124).

These criticisms of CTT boil down to a fundamental problem: the meaningfulness of measures is *relative*; the invariance that we associate with measurement scales for physical properties (length, weight) cannot be achieved in CTT.

3.2.2 Item Response Theory

This problem was addressed by Thurstone as long ago as the 1920's (Thurstone 1959 includes papers from this period) and by Guttman (1950), although IRT is associated more with the names of Rasch (1960), Lord and Birnbaum (Lord & Novick 1968).

Item Response Theory offers techniques for constructing an invariant measurement scale, making *objective* or *fundamental* measurement of psychological traits possible (but not guaranteeing success).

> ... with the careful application of Rasch models, and by invoking the knowledge available for constructing sound tests and questionnaires, it is possible to *attempt* to construct measurements of a fundamental kind in standard test and questionnaire exercises.
> (Andrich 1988:16, italics in original)

Andrich (1988) provides an introduction to the 'sophisticated concept' of fundamental measurement, which, he explains, 'in its most elementary form ... simply allows for arithmetic operations of addition and subtraction on measures' (Andrich 1988:17). The following presentation outlines the basic principles of IRT, without attempting a technical demonstration of how IRT achieves fundamental measurement.

Imagine a group of persons and a group of test items. Assume that the items test a single ability, but that initially nothing is known about the ability of particular persons or the

difficulty of the items. When the persons respond to the items, we learn enough about the two groups to be able to rank them: comparing persons, some are more able than others; comparing items, some are more difficult than others. It is also possible to compare items with persons: a person who answers an item correctly is in some sense *better than* the item, while a wrong answer suggests the person is *not up to the level* of the item. Assume *transitivity* of item difficulty (i.e. that if a person can answer a given item he will also be able to answer an easier item; but if he cannot answer an item he will also be unable to answer a more difficult item). Then comparisons of the kind *rather too easy*, *much too difficult* can be made.

What this shows is that persons and items can both be located on a single continuum which simultaneously describes both ability and difficulty. This is the *latent trait*, or hidden dimension, along which both items and persons can be ranged. This scale is sketched in Figure 3.1.

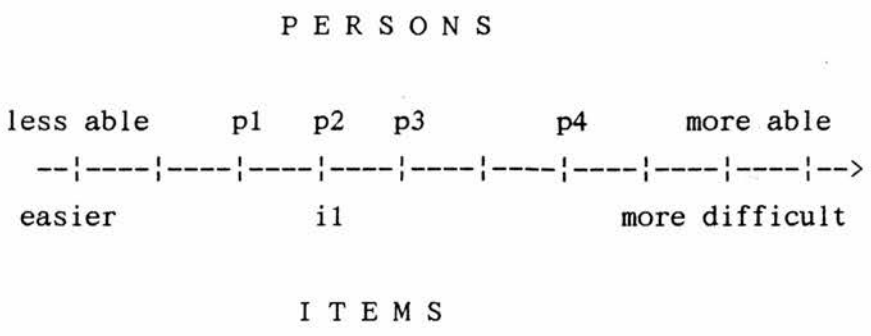


Fig. 3.1: The ability-difficulty scale

The figure shows four persons (p1 - p4) located on the scale according to their ability, and one item (i1) located according to its difficulty. Without knowing the *actual* score by each person on item i1 on the scale, we could hazard the guess that person 1 answered wrongly, while persons 3 and 4 answered correctly. Person 4 might of course answer wrongly, but if he did we would be more surprised than if person 3 did, because

person 4 is much further above i1 on the scale than is person 3. Person 2, who is at the same level as the item, clearly has a 50 per cent chance of answering correctly.

This shows that the *probability* of a particular response by a person to an item is a function of the relationship between the ability of the person and the difficulty of the item. An IRT model is simply a mathematical function which relates these three things. The simplest IRT model - the Rasch model - can be expressed in the following form:

$$P_{n,i} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)}$$

where $P_{n,i}$ = is the probability of a correct response by person n on item i, B_n is the ability of person n, and D_i is the difficulty of item i. The following points about the equation are worth understanding:

$\exp(B_n - D_i)$, which means the exponent of $B_n - D_i$, represents an underlying multiplicative relationship between ability and difficulty, cast into a convenient additive form through a logarithmic transformation.

The value of $\exp(B_n - D_i)$ tends towards 0 as D becomes greater than B, and towards infinity as B becomes greater than D; thus probability of a correct response $P_{n,i}$ moves within the range 0 to 1.

When $B_n - D_i = 0$, $\exp(B_n - D_i) = 1$, hence $P_{n,i} = .5$, i.e. when ability is the same as difficulty, there is a 50% probability of a correct response.

Probability is never exactly 0 or 1, as there is always a small chance of an unexpected correct or incorrect response. The shape of the curve relating ability and difficulty to probability of

correct response is shown in Figure 3.2. The slope of the curve is steepest at B - D = 0, which means that the most information about a person is obtained from an item at exactly the same location on the scale.

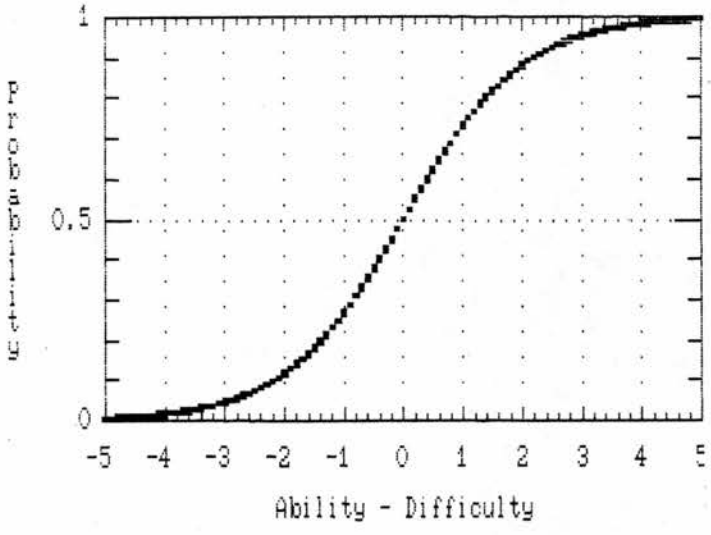


Fig. 3.2 The relation of ability-difficulty to the probability of a correct response in the Rasch model

While ability and difficulty are parameters whose values cannot be known exactly, they can be estimated from the proportion-correct totals for items and persons in a trial test. Obviously chance plays a part in answering an item correctly, so that theoretically *any* set of values for abilities and difficulties might produce the observed set of responses. But some are much more likely to do so than others. Estimation typically involves finding the set of ability and difficulty values *most likely* to produce the observed pattern of responses.

As IRT estimates of ability and difficulty allow precise quantification of the probability that a person of given ability will correctly answer an item of given difficulty, it is possible to compare the *actual* performance of each item and each person with the *predicted* outcome, and thus to obtain a measure of the adequacy of the test as a whole, and of the individual items. Where differences between actual and predicted outcomes are small, the data are said to *fit the model*. This is one important condition for establishing the reliability and validity of a test. Where differences are large - that is, when there is a tendency for persons and items to perform unpredictably - then the model's assumptions must have been violated in some way, and the scale cannot be held to depict a single dimension of ability or difficulty.

As probability of correct response is a function of a simple subtraction involving person ability and item difficulty, the scale is linear. Because no test can show that a person knows absolutely nothing or absolutely everything about a subject, the scale has no end points. In fact, given that ability and difficulty are relative notions, and defined in terms of each other, there is no basis for relating the scale to any fixed point at all. The origin (centre) of the scale must thus be arbitrarily set, customarily at the point of mean difficulty of all the calibrated items. This might suggest that IRT's ability-difficulty estimates are just as sample-dependent as the analagous measures in classical test theory. But there is an important difference. The IRT scale, being linear and invariant, allows items measuring the same trait and persons belonging to the same general population to be fitted subsequently onto the same scale, after suitable adjustment for that first arbitrary centring. The trait constitutes a domain within which any selection of items will produce the same estimate of ability.

3.2.3 Advantages claimed for IRT

Henning (1984, 1987) reviews the advantages of latent trait methods (IRT). The chief advantage claimed for IRT is that measurement of both items and persons is *sample-free*. Within the universe defined by the trait, estimates of ability do not depend on which items are used. Conversely, estimates of item difficulty do not depend on which persons they are trialled upon.

The other major advantage of IRT is that a separate estimate of error is made for each item and person, rather than the one global estimate of test reliability obtainable by classical methods.

While classical test theory routinely analyses item discrimination, it is rare for similar attention to be paid to the performance of persons. The pattern of responses that produce a given score is not examined. And yet an unusual response pattern gives important information about the person. He may be guessing, or not cooperating, or might simply find different things easy or difficult - in which case the construct which the test purports to measure is invalidated, at least for this person. IRT takes into account the way both items and persons perform.

A *norm-referenced* test relates an individual's performance to that of a representative group. A *criterion-referenced* test, in contrast, assesses performance in terms of mastery of certain clearly-stated objectives of learning. The latter approach emphasizes the importance of relevant, useful test tasks, while the former attempts to ensure reliability of measurement by maximizing variance - spreading candidates out along the scale. Henning (1987:111) claims that IRT can reconcile these divergent approaches, one test being capable of both norm-referenced and criterion-referenced interpretation. This is because an ability estimate simultaneously relates a person to other persons

(norm-referencing) and to his probable performance on any items (criterion referencing). Rating a range of criterion tasks in terms of relative difficulty is an attractive possibility for IRT.

IRT facilitates item banking by allowing all items to be placed on a common scale. In fact without IRT it is hard to see how item banking could be done. The advantages claimed below for item banking are thus advantages of IRT.

Item banking proceeds by adding new items to those already in the bank. By using a small link of already-calibrated items in tests of new items one can adjust the ability and difficulty values calculated to fit them to the single scale. In this way items calibrated in different test administrations can be added successively to the item bank, so that it can grow to contain a large number of items, covering the whole range of abilities encountered among learners. Item banking is thus a special case of test equating. Lord (1980:194) says: 'Item response theory is the only method that can carry out vertical equating effectively.' Through item banking learners of differing ability can be given different tests, appropriate to their level, and the results can be directly compared.

Moreover, as more and more items are added to the bank, the picture that emerges can be treated as an operational definition of the ability being measured. Studying what makes items difficult or easy seems a promising way of validating the construct that the bank purports to measure. (see section 3.3.3 below). Item banking thus offers both great practical advantages, and theoretical insights into the nature of the ability tested.

3.2.4 Assumptions of IRT

The advantages offered by IRT outlined in the previous section can only be obtained if the test data to which the model is applied satisfy certain assumptions.

The most important assumption which all practical IRT models work on is that items have an intrinsic level of difficulty and that persons have an intrinsic level of ability on the trait measured. A test of general knowledge including, say, items on pop music and architecture would not fit an IRT model for the obvious reason that certain questions would be simultaneously easy for some persons and difficult for others (depending on their interests). The test would not be measuring one coherent trait: it would not be *unidimensional*. Unidimensionality will be discussed at greater length below (part 3.3.2) as an important aspect of construct validation; suffice it to recall here that 'unidimensionality is a relative matter' (Andrich 1988:9). Thurstone (quoted in the introduction to Chapter 2) recognized that unidimensionality must be *imposed* on a trait by the construction of a suitable testing instrument, that is, by the selection of test items that function in a sufficiently coherent way.

IRT assumes that items are independent of each other, in the sense that performance on one should not be predictable from the answer to any other one (at a given level). A cloze test where answering an item is a help in answering the next one would violate this assumption.

Local independence is in fact a condition of unidimensionality. To understand why this is so, consider the example of the general knowledge quiz above. A person's first responses to questions on pop music and architecture would probably soon reveal that the person knew more about the one than the other. Knowing this, one could make a better prediction as to how he would fare on

subsequent questions. Thus the violation of the unidimensionality assumption is equivalent to a violation of local independence.

Tests must not be speeded. This is particularly vital in tests used for item calibration, as the observed difficulty of an item would be related to its position early or late in the test, and not to its intrinsic difficulty.

3.2.5 Choosing an IRT model

An IRT model describes in precise mathematical terms the relationship of hypothetical constructs (ability, difficulty). The model is an idealization. Real test data tends not to behave ideally. Thus when the model is applied to real data (when values are estimated for the variables, or parameters, in the model) there will be a degree of mismatch between predicted and actual outcomes. This mismatch can be reduced either by making the data behave more ideally (by improving the items), or by modifying the model, probably by making it more elaborate.

Many IRT models have been proposed; but three in particular have been widely applied and investigated, and it is these that will be introduced briefly here. They are the one-parameter logistic (Rasch) model (Rasch 1960, Wright & Stone 1979), the Birnbaum two-parameter logistic model (Lord & Novick 1968), and the Birnbaum three-parameter logistic model (Lord and Novick 1968). The models are related, differing in the number of measurement parameters they incorporate, and thus in the degree to which they accomodate real-world data. The Rasch model has only the single ability-difficulty parameter. The two-parameter model adds a parameter for item discriminability, and the three-parameter model adds to this a parameter for guessing.

The arguments in favour of particular models concern accuracy of estimation, practical and economic considerations, but perhaps most importantly, a fundamental difference in philosophical orientation. While there may be compelling reasons for selecting a model for a particular application *before* collecting data, it has been recommended that different models should be tried out on collected data and the one chosen which performs best (Hambleton & Swaminathan 1985).

The more elaborate three-parameter model can be expected to perform better where all the parameters are necessary to explain the data: that is, in cases where items vary greatly in discrimination, and guessing is a factor in accounting for scores. In other cases a simpler model may work as well or better. The main users of the three-parameter model appear to be large testing corporations and educational bodies who have access to large amounts of data, as well as an interest in maintaining the viability of rather old-fashioned tests that use the multiple-choice format.

With small samples of, say, 100 persons the Rasch model may be the only possible choice. The two-parameter model requires 200 to 400, and the three-parameter model 1000 to 2000 persons for parameter estimation to be accurate. It follows that the more elaborate models require bigger computers and longer run times, and are thus more expensive to use.

The fundamental difference in philosophical orientation concerns an argument whether the model should be made to fit the data, or the data made to fit the model. Proponents of the one-parameter or Rasch model claim that *only* this model can achieve objective measurement, and that the more complex models only 'work' by imposing arbitrary constraints on the values that parameters are allowed to take in the estimation process. Wright, possibly the most prominent advocate of the Rasch model, describes it thus:

> The Rasch model is not a data model at all. You may use it with data, but it's not a data model. The Rasch model is a definition of measurement, a law of measurement. Indeed it's *the* law of measurement.... The Rasch model is ... our guide to data good enough to make measures from.
> (Wright 1988:7)

In other words, if test items do not fit the model, there is something wrong with the items, not the model.

Proponents of the contrary position find the Rasch model over-simple in its assumptions that items discriminate equally and that guessing is not a factor in test performance.

> These assumptions about items fly in the face of common sense and a wealth of empirical evidence accumulated over the last eighty years. Common sense rules against the supposition that guessing plays no part in the process for answering multiple-choice items. This supposition is false, and no amount of pretense will make it true. The wealth of empirical evidence that has been accumulated concerns item discrimination. The fact that otherwise acceptable achievement items differ in the degree to which they correlate with the underlying trait has been observed so very often that we should expect this kind of variation for any set of achievement items we choose to study.
> (Traub 1983:64)

Advocates of the Rasch model sometimes appear to occupy a rather paradoxical position, as together with strong statements such as the above, they readily admit that real data *never* fit the model perfectly. The question is, then, how one is to decide whether or not the data is 'good enough to make measures from.' As Hambleton & Swaminathan (1985:155) note, 'there is evidence that the [IRT] models are robust to some departures, but the extent of robustness of the models has not been firmly established.'

Not all advocates of the Rasch model argue for its theoretical superiority. Henning (1987:116), for example, advocates the one-parameter model for 'utilitarian decision-making purposes' where it does not greatly matter if rather more items are rejected as misfitting.

There are two practical reasons for using the Rasch model for this study. Firstly, sample sizes will be limited, and secondly the whole item bank, including the software for parameter estimation, is implemented on a microcomputer.

The argument that the Rasch model is more objective, though supported with mathematical proofs which are difficult for the non-mathematician to follow, is also attractive. But using the Rasch model means taking particular care with the construction of items, and particularly, eliminating the factor of guessing (Pollitt, personal communication). Given that the purpose of this study is not 'utilitarian decision-making', but rather the validation of a construct, care must be taken in the treatment of misfit. The wholesale rejection of misfitting items is *not* a satisfactory way to construct a unidimensional trait which is to be theoretically interpretable.

In its simplest form an IRT model deals with dichotomous data (items marked right or wrong); but IRT models can be extended to deal with various forms of partial credit or rating scale (Wright & Masters 1982); one 'multi-faceted' extension of the Rasch model treats item difficulty and person ability as just two of many possible factors relating to test performance (Linacre 1989). The present item bank uses dichotomously-marked items.

3.3 Construct validation and IRT

3.3.1 Construct validation

Bachman (1990) provides an interesting discussion of the concept of validity. He first discusses the relation of *reliability* to validity:

> We might say that reliability is concerned with determining how much of the variance in test scores is reliable variance, while validity is concerned with determining what abilities contribute to this reliable variance (Bachman 1990:239).

He quotes this classic statement of the relationship between reliability and validity by Campbell and Fiske (1959):

> Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods. (Campbell and Fiske 1959:83)

Bachman points out that in language testing the two concepts can be hard to distinguish clearly, given that the distinctiveness of test methods is frequently not clear.

> For example, is the correlation between concurrent scores on two cloze tests based on different passages to be interpreted as reliability or validity? It depends upon how important we believe (or know, from actually examining this facet) the effect of the difference in passages to be. (p.240)

Furthermore, given that in language testing, 'what is trait and what is method is very hard to distinguish' (Stevenson 1981:53), the distinction drawn by Campbell and Fiske can be hard to apply. 'No method it seems to me can ever be entirely free of the trait it seeks to realise' (Davies 1981:184).

Bachman presents a view of validity as a unitary concept, subsuming the various types of validity traditionally identified, such as content, criterion and construct validity.

> Validity ... is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (American Psychological Association 1985:9)

Bachman defines *construct validity* as 'the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs,' and presents construct validity as a unifying concept which subsumes both content and criterion validity.

> Virtually all test use inevitably involves the interpretation of test scores as indicators of ability, and as soon as we ask the question, 'What does this test really measure?' construct validation is called for. Construct validity is thus seen as a unifying concept, and construct validation as a process that incorporates all the evidential bases for validity discussed thus far. (Bachman 1990:256)

There are many types of empirical evidence that can be used in the process of construct validation. The following discussion considers the analysis of data-model fit in IRT as an approach to construct validation. Then the investigation of item difficulty,

which also follows naturally from adopting an IRT orientation, will be presented as a further important aspect of construct validation.

3.3.2 Unidimensionality

The analysis of data-model fit in IRT relates to how well a set of test items delineate a *unidimensional* trait. The notion of unidimensionality is one that we should now examine in more detail.

In a certain sense it is obvious that any good test must be unidimensional. 'Measurement is essentially a one-dimensional process' (Choppin 1981:205). Yet the appropriacy of IRT as a model for measuring educational achievement has been challenged, precisely because of its insistence on unidimensionality. The argument that unidimensionality is not *desirable* merges with an argument that it is impossible to achieve, for any serious measurement purpose. These arguments are clearly relevant to a discussion of validity.

Goldstein (1981) argues that the Rasch model is inappropriate for dealing with the necessarily varied aims and methods of teaching. Tall (1981:192) calls the model 'in educational terms unbelievably naive', and warns that the logic of the Rasch model inhibits new developments in teaching by penalizing innovation (though it might be argued that *any* established test tends to under-value innovations in teaching). Traub (1983) states:

> It will be a sad day when our conception of measurable educational achievement narrows to the point where it coincides with the criterion of fit to a unidimensional item response model, regardless of which model is being fitted. (Traub 1983:19)

Goldstein (1981) and Tall (1981) both attack the assumption that test items have an intrinsic difficulty which remains stable over time (a criticism particularly directed at item banking). Tasks may become more difficult over time, as their content becomes dated, or easier, as teaching methods change and develop. Goldstein argues that selection of items simply because they fit the model leads to unrepresentative test content.

It is not clear that these criticisms of IRT are all relevant to the measurement of language proficiency. The investigation of language proficiency in Chapter 2 found evidence that it *is* a reasonably homogeneous construct, however measured. Ebel (1979) cites foreign languages and mathematics as subjects whose homogeneity makes high reliability easier to achieve than in some other tests of educational achievement. Lord (1980:190), although he does not mention foreign language learning as such, lists vocabulary, reading comprehension and verbal reasoning with subjects that are 'likely to be reasonably unidimensional'.

But Goldstein's argument concerning the impermanence of tests certainly applies to language tests too. Language proficiency exams of today test certain skills perhaps not considered relevant thirty years ago. Standards of assessment also change: communicative effectiveness is presently more highly rated, and errors of accuracy less heavily penalized, than was true not so many years ago (Anthony Howatt, personal communication). An item bank for measuring language proficiency cannot be made immune to the fact that perceptions of what is a relevant test of proficiency may change as theories of learning change. In any case, the question of preserving unidimensionality over time cannot be investigated by the present study, and so will not be considered further.

As to the general charge that the 'unidimensionality assumption' will always be violated in the case of measures of complex educational achievements, this may represent a misunderstanding of what is meant by 'unidimensional', or by 'assumption'

(McNamara 1990:106); as we have already seen, the IRT view is that the imposition of unidimensionality is a necessary condition of measurement, and that the simplification of reality which it entails is (or may be) compensated for by the utility of the results.

> Unidimensionality is a relative matter - every human performance, action, or belief is complex and involves a multitude of component abilities, interests and so on. Nevertheless, there are circumstances in which it is considered useful to think of concepts in unidimensional terms.
> Andrich (1988:9)

As Lord & Novick point out (1968:358), a trait orientation to psychological theory does not necessarily imply that traits exist in any physical or psychological sense. We have already cited Hamp-Lyons (1989:115):

> Unidimensionality ... is a psychometric property independent of any concept of 'dimensions' of language proficiency, which are psycholinguistic properties or concepts.

McNamara (1990:112) stresses the importance of distinguishing between measurement models, and models which make explicit claims about the various skills and abilities underlying test performance.

> The measurement model posited and tested by IRT analysis deals with the question, 'Does it make sense in measurement terms to sum scores on different parts of the test? Can all items be summed meaningfully? Are all candidates being measured in the same terms?' This is the 'unidimensionality' assumption; the alternative position requires us to say that separate, qualitative statements about performance on each test item, and of each candidate,

> are the only valid basis for reporting test performance.

McNamara discusses a study by Henning, Hudson & Turner (1985). Henning *et al* examined the UCLA English as a Second Language Placement Examination (ESLPE). The test consists of 150 multiple-choice items, 30 in each of five sub-tests: Listening Comprehension, Reading Comprehension, Grammar Accuracy, Vocabulary Recognition and Writing Error Detection. The performance of 300 learners of various language backgrounds was examined. Although 11 items were identified as misfitting, Henning *et al* concluded that they had satisfactorily defined a single dimension of ability and difficulty, and hence that Rasch analysis was

> sufficiently robust with regard to the assumption of unidimensionality to permit applications to the development and analysis of language tests (Henning *et al* 1985:152).

As McNamara notes, the 'robustness' which Henning *et al* here take to be a virtue is from another point of view worrying, as

> the unidimensional construct defined by the test analysis seems in some sense to be at odds with the *a priori* construct validity, or at least the face validity, of the test being analysed (McNamara 1990:110).

Hamp-Lyons (1989) cites the Henning *et al* study in her critical review of an article by Adams, Griffin & Martin (1987). This is one of a series of reports on the development of the Interview Test of English as a Second Language (ITESL): Griffin, Adams, Martin & Tomlinson (1986), Adams, Griffin & Martin (1987), Griffin, Adams, Martin & Tomlinson (1988). Adams *et al* (1987:24) describe work that shows the latent trait approach to be 'successful in defining a grammatical or structural dimension'. They conclude more generally 'that the Rasch model can be used as part of a confirmatory approach to dimensionality studies' (p.24). Hamp-Lyons takes this to be a strong claim about the

underlying components and processes of language development. She objects (citing the Henning *et al* study) that Rasch analysis is too tolerant of violations of unidimensionality to be of use in demonstrating the reality of a posited 'grammatical dimension'. Nunan expresses the same criticism more strongly:

> [The test] illustrates quite nicely the dangers of attempting to generate models of second language acquisition by running theoretically unmotivated data from poorly conceptualized tests through a powerful statistical programme.
> (Nunan 1987:156 quoted in McNamara 1990:115)

As McNamara says in defence of Adams *et al*, their intention is not to 'generate models of second language acquisition': Hamp-Lyons and Nunan are, he says, failing to distinguish between the two types of model.

Henning *et al*'s findings that the ESLPE test fits a single dimension need not actually be taken as evidence that Rasch analysis is overtolerant of violations of unidimensionality. All parts of the test being multiple-choice, we can expect method at least to obscure trait. Also we have seen that there is a large general factor across tests of putatively different language skills. Thus the result is not surprising. In any case, Henning reports that factor analysis also produced a single factor solution for the same test.

The question is perhaps not whether Rasch analysis is refined enough to detect violations of unidimensionality; that must depend more on the nature of the data. As Andrich (1988:62) points out:

> ...the chance a model will fit varies inversely with the precision of the estimates. Therefore, the poorer the precision the more likely data will fit the model. Furthermore, the precision of the estimates depends upon the

> sample size - the larger the sample size, the greater the precision. Thus the larger the sample, the less likely the data will show fit to the model.

More to the point perhaps is that the orientation of Rasch analysis tends to be more towards demonstrating fit than detecting misfit. As Linacre (personal communication) puts it:

> The Rasch model is not a *descriptive* model, but a *measurement* model. The essential question is 'are the data a good enough fit for the measures to have generalizable meaning?' ... In this respect, Rasch analysis is closer to meta-analysis (research synthesis), with its emphasis on effect size, than to the conventional analysis of descriptive statistical models with its emphasis on hypothesis testing. (emphasis Linacre's)

In fact, McNamara's (1990:107) distinction between the 'measurement dimension which is *constructed* by the analysis' (the IRT trait) and the 'dimensions of underlying knowledge or ability which may be hypothesized on other, theoretical grounds', is a familiar one in testing and construct validation in general. The fact that the distinction seems frequently to be lost sight of in discussions of IRT may have something to do with the more unfamiliar 'dimension' metaphor, or perhaps with the seductive power of 'powerful statistical programmes'. Classical approaches to construct validation, such as predicting group differences, or looking at internal correlation of items in subtests of a battery, are essentially based on demonstrating the unidimensional behaviour of items. This unidimensionality has two aspects: it associates diverse items that measure the same construct, and distinguishes items that measure different constructs.

Studies of fit in IRT are called construct validation through internal correlation by Henning (1987:115). But in classical approaches as in IRT, the empirical demonstration of

unidimensionality is insufficient on its own to establish construct validity, even if some writers appear to suggest otherwise. Griffin *et al* (1988:9) write of the ITESL test:

> Following Wright and Masters' (1982) definition, construct validity was taken to mean the existence of a measurable dimension in the test.

As is clear from Griffin *et al*'s discussion, the empirical demonstration of a 'measurable dimension' is taken to support the construct validity because it confirms an *a priori* theory. Construct validity means not only *constructing* a unidimensional trait, but *interpreting* it in the light of theory.

> The response model does not embody a construct theory. The fit of data to a particular response model can be tested without knowledge of where the data came from, whether the objects are people, nations or laboratory animals, or whether the indicants (i.e., items) are bar presses, scored responses to verbal analogy questions, or attitudinal responses. Nothing in the fit between response model and observation contributes to an understanding of what the regularity means. In this sense, the response model is atheoretical. Once a set of observations has been shown to fit a response model, the important task remains of ascribing meaning to scaled responses. In a way, the distinction is similar to that between classical reliability and validity. Like a well-fitting response model, high reliability suggests that 'something' is being measured; but what that 'something' is remains to be specified. (Stenner, Smith & Burdick (1984:308)

Hambleton & Swaminathan concur:

> The fact that a set of test items fits one of the item response models indicates that the items measure a common trait and nothing more. What is needed is a construct validity study to determine the characteristic(s) or trait measured by the test.
> (Hambleton & Swaminathan 1985:70)

They add that content validity studies, while important, are probably not sufficient.

Good fit suggests that something has been measured; but determining what depends on interpreting the constructed trait in terms of theory.

There are warnings in the IRT literature against over-interpreting random effects. Wood's (1978) description of data from a coin-tossing experiment that fitted a response model well is frequently cited in this context. It is also pointed out that the practice of achieving good fit by mechanically rejecting all badly-fitting items may change the characteristics of the item domain in '(perhaps) subtle or unknown ways' (Hambleton & Swaminathan 1985. See too Goldstein 1981).

Divergence from perfect fit can be treated analogously to the case of subtests in a test battery. Choppin (1981) defends the Rasch model from accusations of over-simplicity, arguing that the very existence of divergence from perfect model fit offers useful insights into the nature of the constructs tested. He proposes two approaches to dealing with divergence.

First, a number of traits may be identified within the overall trait, and the learner's proficiency reported as a profile, not a simple score. We might find, for example, that the proposed trait 'linguistic competence' might be better divided into 'vocabulary' and 'grammatical competence'.

Secondly, (and preferably) explicit use is made of deviations from the simple measurement model. Choppin proposes to calibrate every item in terms of its difficulty considered as an indicator of achievement in the general domain, 'and also as regards the information it carries in terms of its deviation in a particular direction.' Gathering together performance on one sub-domain should provide 'an indication (even a measure)' of the extent this departs from performance in general. Thus, for example, the article system might be found to present a certain level of difficulty overall, but to be more difficult for Poles than Italians. As Goldstein (1979) points out, the analysis of divergence from the model only makes sense if the model itself is valid, which is the assumption Choppin is making.

Investigating model fit thus contributes to construct validation, on condition that there is a construct to validate: that is, performance of items is interpretable in terms of theory, and fit is not achieved simply by rejecting without investigation all ill-fitting items.

3.3.3 Explaining item difficulty

Stenner, Smith & Burdick state that the study of variation in question difficulty is the most promising way of understanding the construct measured by the test.

> Until the developers of educational and psychological instruments can adequately explain variation in item scale values (i.e. item difficulty), the understanding of what is being measured will remain unsatisfyingly primitive. (Stenner, Smith & Burdick 1984:305)

Pollitt says:

> Traditional approaches to construct validity have been indirect, looking at the abilities of students as revealed by tests rather than directly at the tests themselves. The essence of item banking is that the questions *are* the scale. The construct 'Listening ability' is defined by the relative ordering of the questions in a listening bank; the factors or characteristics of test questions that determine their relative difficulties also determine just what it is that the test truly measures.
> (Pollitt 1990:878)

The important question is, as Pollitt & Hutchinson (1986) ask: 'What makes questions difficult?' By *explaining* the relative difficulties of items, one is explaining what it is that the test measures.

Stenner, Smith & Burdick propose four advantages of focusing on variation in item scale values (pp.309-310):

1. Stating theories so that falsification is possible.

2. Item scale values are typically more generalizable, i.e. reliable, than are person scores (because they are derived from the answers of large numbers of persons)

3. Ease of experimental manipulation: 'Items are docile and pliable subjects.'

4. Intentions can be explicitly tested.

Pollitt & Hutchinson (1986) claim additionally:

5. Questions are of central concern in test construction. Validity can be built into the construction process by using an explicit theory.

6. A test is a set of questions, not of people. 'If we wish to validate a test it seems more obvious to seek to understand the behaviour of its questions rather than the behaviour of some people.'

Stenner, Smith & Burdick propose a *construct specification equation* to explain observed variation in item scale values. In the example they present, receptive knowledge of vocabulary is predicted from a construct specification equation involving three factors: the common logarithm of the frequency of a word's appearance in large samples of written material; its dispersion over a range of texts, indicated by a value from 0 to 1, and its abstractness - a dichotomous feature assigned by judges. A response model like the Rasch model allows observations to be compared with predictions, so that the specification equation can be verified or modified. Thus the difficulty of receptive vocabulary test items is explained as a function of word frequency, range and abstractness, the relation being stated in explicit mathematical terms.

Pollitt & Hutchinson, while pursuing a similar line of enquiry, have this criticism:

> The American approach is strongly empirical, concerned with finding question characteristics that maximise the predictive power of the model, and so runs a risk of imputing causation to mere correlations.
> (Pollitt & Hutchinson 1986:43)

They propose instead 'an explicit model of the answering process.'

Pollitt & Hutchinson attempted to quantify the factors identified as adding to or removing difficulty from tests of English, French, geography, mathematics and chemistry for 16 year-old children, and were able to pursue the attempt furthest for tests of reading comprehension.

First a detailed scheme was devised, modelling the answering process: reading the question, searching the text, understanding the meaning of the relevant parts, and composing a response. This scheme was 'only partly based on evidence', 'plausible rather than proven', and played a heuristic role. It attempted to be exhaustive to provide plenty of variables for analysis.

A multiple regression analysis was used, implementing a causal model in which the answering process is strictly sequential, and failure at any point leads to a wrong answer. The analysis thus revealed those steps in answering a question that added significantly to difficulty.

The findings are interesting inasmuch as the eight significant variables include potential ambiguity in the question, interaction with an earlier question, and the need to write an answer requiring more than simple quotation - all factors which contradict our ideas of what a reading comprehension test *should* measure. Other factors accord well:

> Successful candidates are those who can avoid being distracted from the real meaning by dominant or emotive words, who can synthesize answers from separate pieces of text and who can cope with difficult words and syntax. These surely are the kinds of abilities we want a reading test to measure.
> (Pollitt & Hutchinson 1986:57)

They conclude:

> Our aim is to combine explicit theory with empirical confirmation; not theory about Reading but theory about Reading Assessment. It is the congruence between what a reading test tests and what we think it should test that constitutes validity.
> (Pollitt & Hutchinson 1986:60)

The Stenner *et al* and the Pollitt & Hutchinson studies are offered as illustrations of construct validation through the explanation of item difficulty. The two studies differ chiefly, perhaps, in the complexity of the construct dealt with: 'receptive vocabulary knowledge' seems a simpler notion than 'ability to answer reading comprehension questions', and thus in the first case the 'explanation' of item difficulty hardly needs to incorporate a statement of causal order - the 'explicit model of the answering process' proposed by Pollitt & Hutchinson. But causal order is certainly relevant to explaining the difficulty of the items in the present study, and will thus be taken up again in Chapter 6, where empirical work on explaining item difficulty is described.

4: The design of the item bank

4.1 ItemBanker in outline

Item Banker is a database, specially written for the purposes of item banking and test construction. Like any database, it allows for the entry and retrieval of data, the data being test items. The organization of Item Banker reflects chiefly the following two goals:

1) that language teachers should be able, without much training, to produce paper-and-pencil tests having the desired difficulty level and content;

2) that the items should support computer-adaptive as well as paper testing. This entails that items should be markable by computer, which imposes a certain rigour on the forms items can take.

When early versions of Item Banker were tried out by teachers, it became clear that offering access to all of Item Banker's facilities made the system confusing to operate, (unnecessarily so, given teachers' limited goal of producing paper tests). It was therefore decided to cater for two classes of user: 'teachers' and 'system users', the former being offered a simpler, fixed sequence of options, and the latter being allowed unlimited access to the bank's facilities. The following description of Item Banker begins by following through the sequence of screens that teachers are offered, as this provides the clearest outline of the way the bank functions.

4.2 Simple operation by teachers

Teachers are guided through the following sequence of screens:

1) the Overview Screen, which lets users browse through the available question types, mainly for familiarization;

2) the Search Screen, where users specify the kind of items they want, according to difficulty, content and type of question;

3) the Viewing Screen, where they can see each item in the set they have selected, and delete ones they do not want;

4) the Printing Screen, which lets users specify a title for the test, and then offers the choice of printing to a disc file, or directly to printer.

A measure of back-tracking is possible: from the viewing screen, users can return to the search screen and make a different selection. Nonetheless, the overall orientation is to progress forwards towards the goal of printing a test paper.

4.2.1 Overview screen

The top part of the overview screen (see Fig. 4.1) summarises the items available (available here meaning calibrated, as generally speaking teachers will only need to use items which have already been calibrated, and can thus be used in tests). Item numbers are summarized for three groups: RECOGNITION, ONE WORD and LONGER.

Item Banker in fact offers six distinct item formats or types:

1) Jumbled words, e.g:

Put / book / the / table / the / on / .

The first word always remains unmoved, at the beginning of the sentence, and the final punctuation mark remains at the end.

2) Jumbled sentences, e.g:

I come from London.
( ) It's a big ugly city.
( ) But I like it.
( ) That's the capital of England.
( ) Perhaps because I grew up there.

3) Matching pairs, e.g:

John _____ early in the morning.
He _____ to work on the bus.
He _____ early on Fridays.
He _____ late watching TV.

goes off / sits up / gets off / gets up

4) Multiple choice

The classical four- or five-choice format. (Though available, this type is not in fact used in the bank of items which is the subject of this study).

5) Gap fill (one word), e.g:

Is __(there)__ a bank near here?

6) Phrase, sentence, e.g:

When is she coming?
Do you know ...(when she is coming)...?

The RECOGNITION group includes the jumbled words, jumbled sentences, matched pairs and multiple choice item types, while ONE WORD and LONGER include the gap fill and Phrase or sentence types respectively.

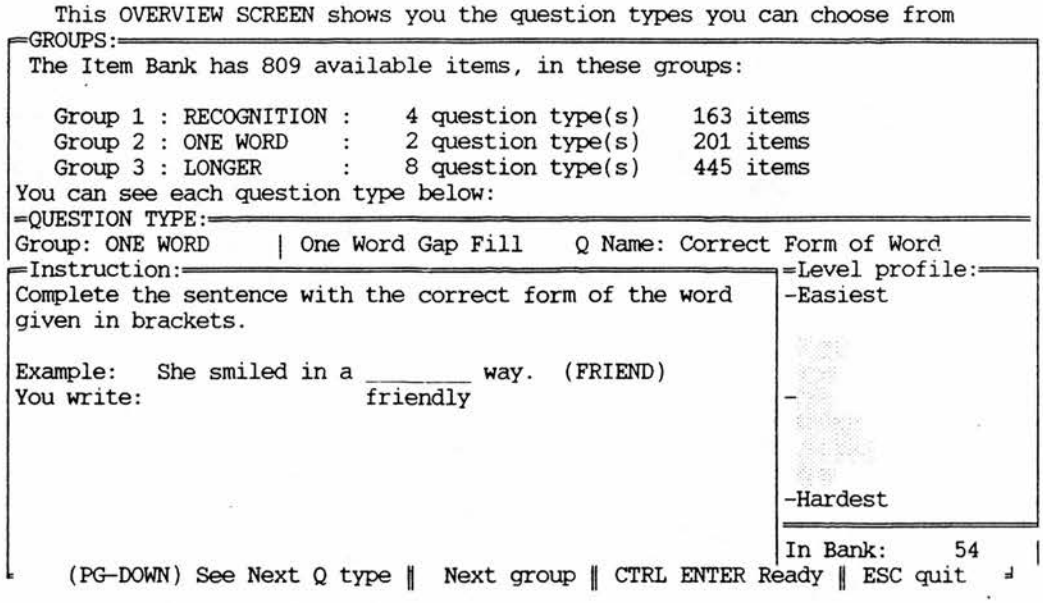


Fig. 4.1. Item Banker: the Overview Screen

It is important to realize that although the number of item types is fixed, each type may support a number of different rubrics, (called 'question types' in the Overview screen). The most productive item type in the present bank is the phrase or sentence type, which supports eight question types, such as:

> Finish the second sentence so it means exactly the same as the first one. ...
>
> Make a question using the word in brackets. ...
>
> Complete the sentence with the correct form of the verb in brackets. ...

The potentially confusing distinction between formally different item types is not made explicit to the teacher, who is only offered choices at the level of 'question type'. The overview screen lets users browse through each group of question types. On

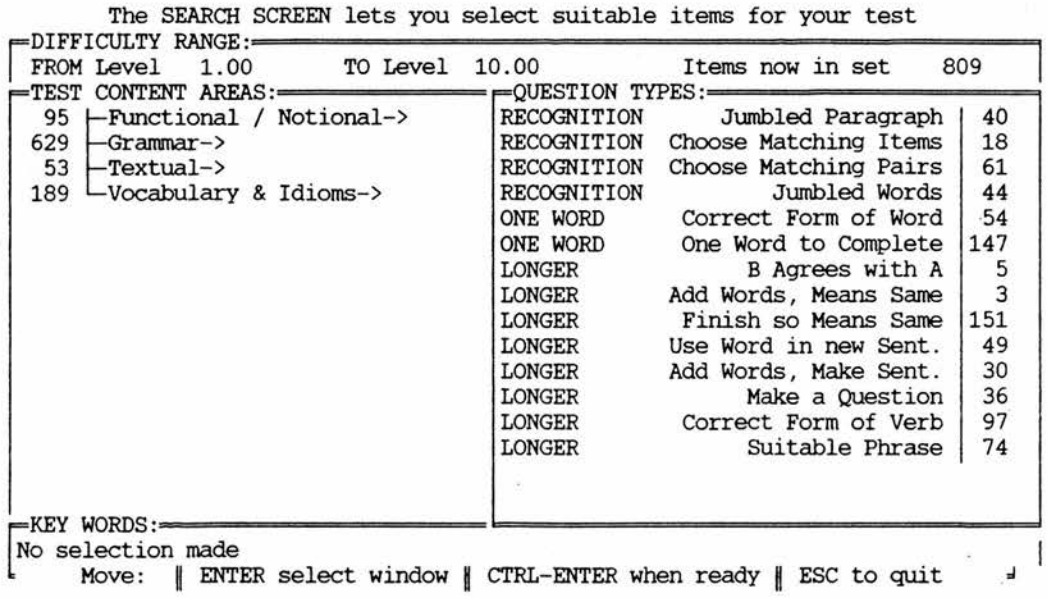
the right, a sideways-on histogram gives a level profile, showing at a glance how the items for each question type are distributed for difficulty through the bank. Below this the number of items for each question type is displayed.

4.2.2 Search screen

From the overview screen teachers move on to the search screen (Fig. 4.2). The screen is laid out to offer choices in the order which, it is believed, corresponds best to teachers' purposes. Thus it is assumed that the teacher starts with a particular group of learners in mind, and so difficulty level is the first choice.

The level is displayed not in logits (which are the units kept in the data record, and used for all statistical operations), but in units of a user-supplied scale. In the illustrations in this chapter, this is a ten-band scale used by Eurocentres. The logit scale is in any case not particularly user-friendly, running as it does from roughly minus four through zero to roughly plus four. It is also important that Item Banker should use a scale which is already familiar, and thus has meaning, within a particular educational setting, if it is to be able to contribute usefully to the overall process of assessment (of course, the equating of proficiency scales derived from different kinds of assessment is itself a difficult question, but it lies outside the scope of the present study, and it will not be dealt with here).

Top-right of screen is the number of items remaining in the current set. As each panel of the screen is edited, this number is updated. Selections can be changed, to adjust the number of items upwards or downwards. Columns of numbers also appear to the sides of the screen, showing (on the left) the number of items on particular content areas, and (on the right) the number of items available for each question type.



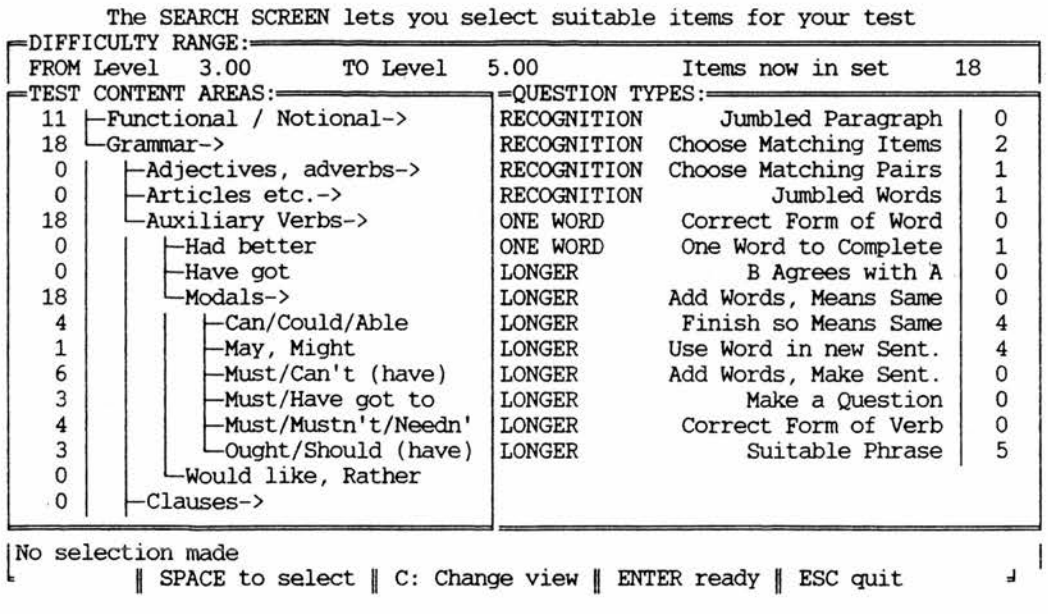


Fig. 4.2. Item Banker: the Search Screen as seen by teachers.
Top: On entry, before making any selection;
Bottom: while selecting content areas.

Next the user can select areas of test content to include. The bank has a set of headings (or 'tags') denoting content areas, and this set is arranged hierarchically, so that smaller areas can be identified within larger ones. This set can of course be extended by a competent user. Each item when added to the bank can be assigned a number of such tags, enabling items to be selected by content. Upon entry to the screen only the most general content areas are shown (in the present bank, there are four of these: Functional/Notional, Grammar, Vocabulary and Idiom, and Textual). When the user selects an area, it is highlighted, and expands to show the sub-headings it contains. The user can then either leave the higher-level, more general heading selected, or select one or more sub-headings (in which case the higher-level selection is turned off). Thus it is possible for test content to be specified with a greater or lesser degree of precision, according to the purposes of the particular test.

This is the great advantage of the hierarchical organization. One disadvantage is that users may not be able to locate a particular content area they have in mind, if they do not know which higher-level heading it lies below. To solve this problem, an alternative alphabetical view of the list can be selected, so that a user looking for, say, 'logical connectors' can check at once if the heading exists.

A refinement to the test content selection process is selection by keyword, which is the next choice offered. Each item record can include one key word or phrase. When the keyword panel is entered, an alphabetical list of all the keywords in the items remaining in the subset is displayed, allowing the user to select or exclude items having particular keywords.

Finally the user selects question types to include. The list only shows short names for each question type, but the whole question type can be viewed upon request. Users will select from

those question types for which sufficient items remain, while also probably trying to keep the number of different question types to a minimum.

4.2.3 Item viewing screen

In the viewing screen (Fig 4.3) users can see each item, and delete it from the set if it is not what they want. Panels to the right of the screen give information about the difficulty of the currently-displayed item, as well as summary information about the set of items: the highest, lowest and average difficulty, and the list of question types used, with the number of items for each question type. Items are displayed sorted into groups by question type, and within each group are further sorted by difficulty, easiest to hardest. This is the order in which they will be printed out in the test.

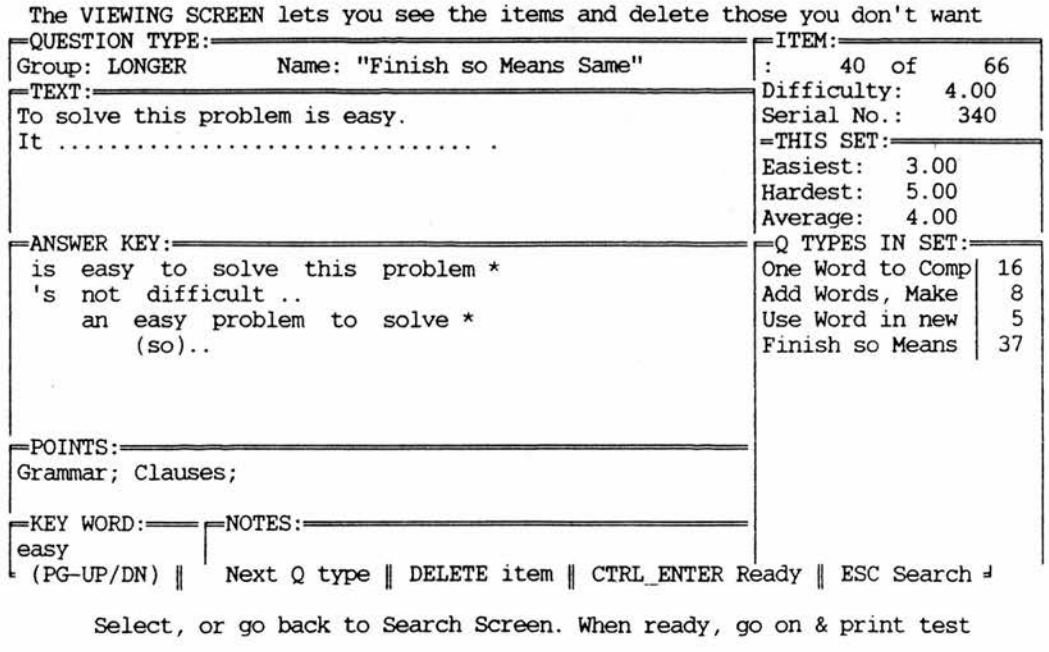


Fig. 4.3. Item Banker: the Viewing Screen as seen by teachers

4.2.4 Printing screen

The printing screen offers few choices to teachers. They can give a title to print on the test, and a file name if the text of the test is to be saved to disc rather than sent to printer. What exactly is printed depends how the system has been set up, but will include the question paper, the answer key, and a transformation table to allow raw scores on the test to be converted directly to band scores.

4.3 Operation by "system users"

In contrast to teachers, system users select what they want to do from menus. The screens described above are also accessible to system users, but with additional options. In the search screen, for example, there are extra menus allowing calibrated or uncalibrated items to be selected, or selection of items falling within certain limits (of Standard Error, t fit or serial number). Extra options in the viewing screen include the possibility of sorting the items in different ways, seeing descriptive statistics on the current set, and so on. The record of tests previously compiled can be consulted, allowing the system user to check what items have been used where. The system user can also see information about each item (its Standard Error, t fit statistic and so on) which is not shown to teachers.

Additionally, system users can do a variety of things not offered to teachers: they can write new items and edit existing ones, add to the sets of available question types or test content areas, input scores from trials of new items, and do Rasch estimation to calibrate items. They can also change certain aspects of the way the bank works for teachers.

4.3.1 Writing and editing items

Writing or editing an item involves selecting the item type, selecting from available question types(or rubrics), and then entering the text of the question. How this is done depends on the item type. Much of the work of arranging items graphically, and jumbling options in the recognition (selected-response) types, is done automatically.

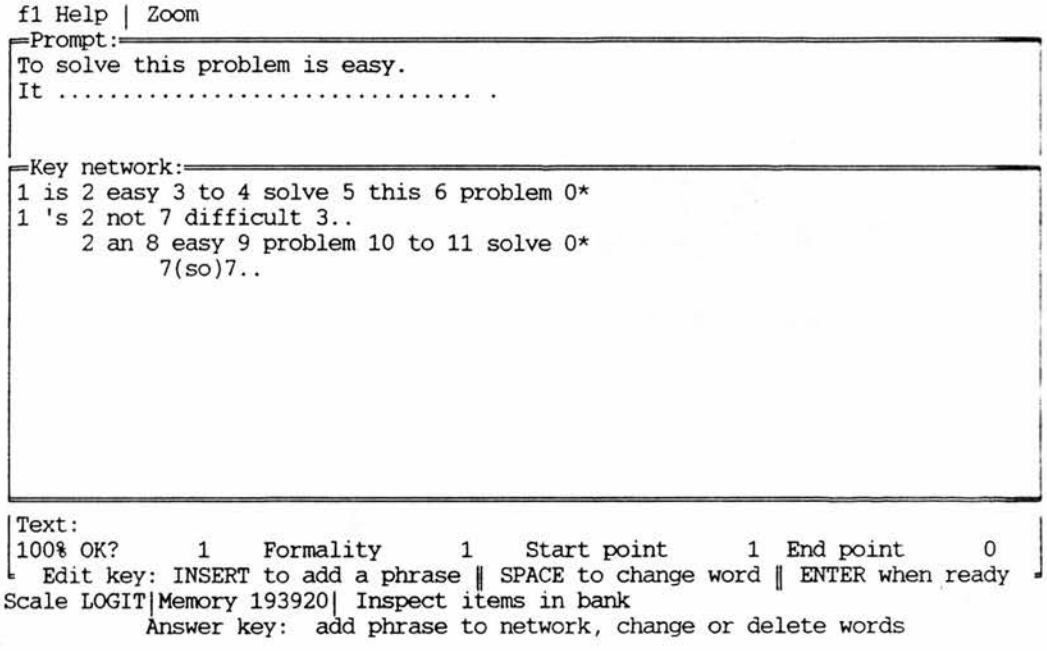


Fig. 4.4. Item Banker: editing an answer key network

An answer key is generated automatically for the recognition types, but for the gap fill and phrase or sentence types all possible answers must be entered by the item writer. A transition network structure is used, which is made to hold a large number of alternative answers in a compact and fairly legible form. Legibility is a potential problem because the answer keys must serve two purposes: in their coded form they are used by the computer-adaptive test to judge responses as right or wrong, while in their printed form they are offered as

answer keys for use by teachers. Fig. 4.3 shows an answer key as it appears in the latter function: legibility is enhanced by showing a minimum of information and relying on a degree of common sense for correct interpretation. Fig 4.4 shows the network in the course of editing and gives rather more idea of its formal structure - a set of numbered nodes linked by arcs: that is, spaces linked by words.

Writing a network is reasonably straightforward. Elements are added a phrase at a time. The computer-adaptive test (CAT) uses fuzzy matching by default - that is, it accepts small spelling mistakes in words of five letters or more. This can be overridden by highlighting letters that must be exactly matched, which is useful where, for example, agreement or verb endings are considered to be important aspects of the item.

Each string of words entered can then be marked as "100% OK", and further characterised as formal, informal or neutral. The "100% OK" flag addresses a problem which inheres in the fact that the bank is likely to be used for two different purposes (this problem is discussed at greater length elsewhere). Many items in the present bank rest on an intended pedagogical point, and yet happen to admit of various other responses which seem acceptable although they do not address this point. For proficiency testing such responses must be judged right, while for some other purposes they are an embarrassment: for example, for a computer-adaptive diagnostic program which uses the bank to provide pedagogic exemplars. Answers not marked as being 100% OK will not be made available for the latter purpose.

Because it operates at the level of single words the network is simple, but at the same time limited in the kind of complexity it can handle (it has no syntactic rules). It makes practical the use of constructed-response item types in computer- adaptive testing, as long as some care is taken in the construction of items. One field in the item record, marked "OK for CAT", can be

set to "NO" if it turns out to be impossible to enter all acceptable responses into the network. Then the item remains available for paper testing, but will not be used by the CAT.

The remaining fields in the item record to be edited include the content points, the keyword, and optional notes on the item. An initial guess at the item's difficulty can be entered, which is useful when selecting new items of similar level for trialling together.

4.3.2 Calibration of new items

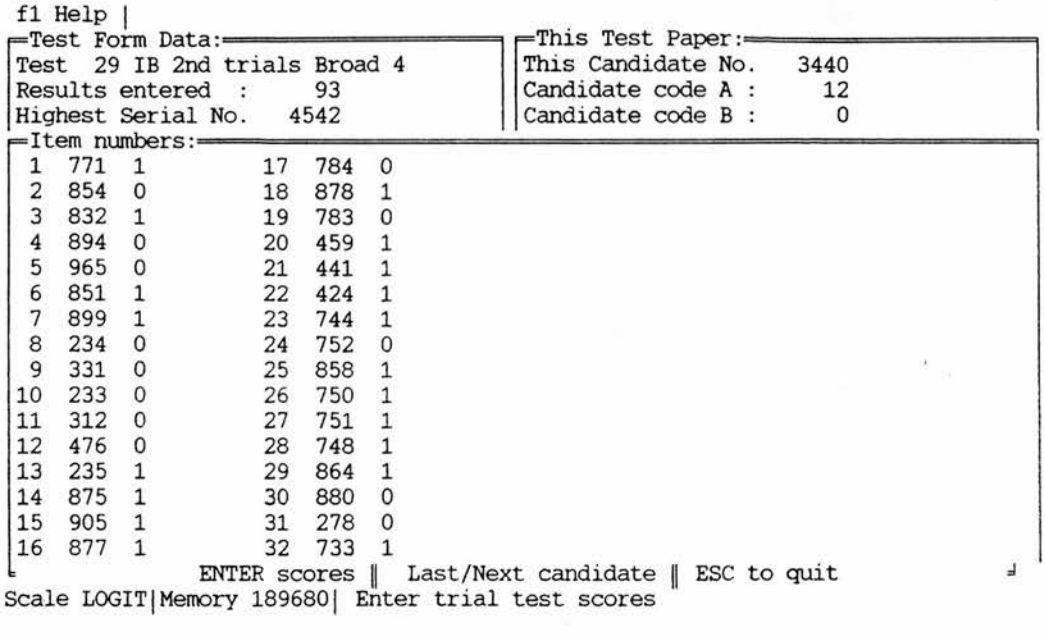


Fig. 4.5. Item Banker: Screen for entering scores

Adding new items to the bank begins with writing the items and trialling them on learners. Scores can then be entered, using the screen shown in fig. 4.5. A test record is selected by name from a list of tests compiled. The screen then shows the items in the test, identified by their number in the test as well as by

their bank serial number. Two fields are available for recording coded information about each candidate; in the present study candidates' L1 was recorded in this way. Scores are entered as 1 or 0, or can be shown as missing (it is an advantage of the Rasch estimation procedure used that missing values are simply ignored, rather than counted as wrong). The input data is saved as a score matrix.

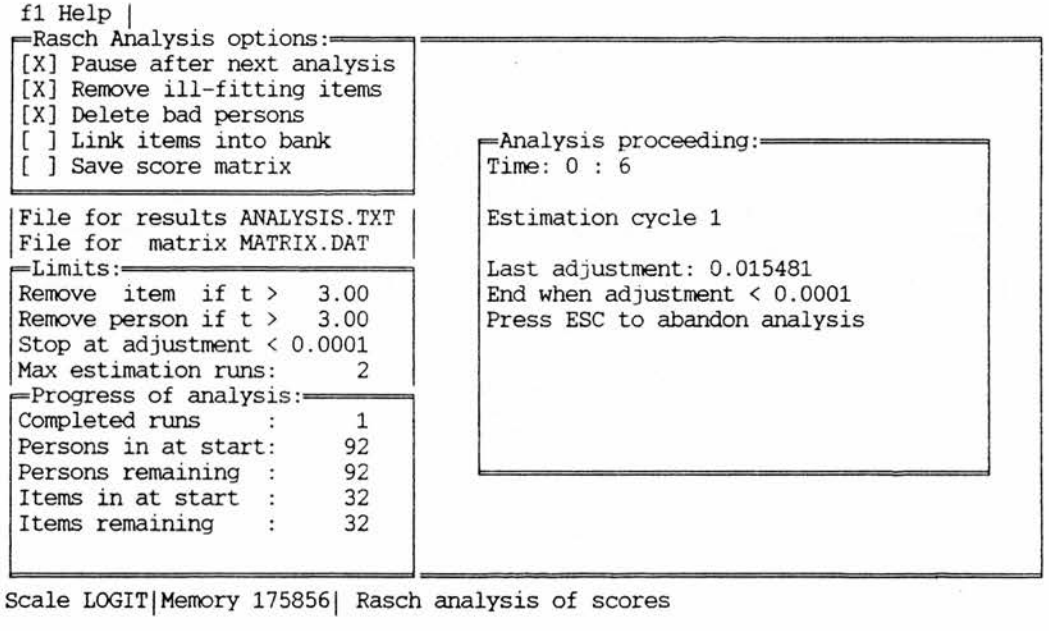


Fig. 4.6. Item Banker: Rasch analysis screen during estimation

The Rasch estimation routine allows the user to select several of these matrices for simultaneous estimation (and performs certain checks to ensure that they contain the necessary common items). There are certain limited possibilities for excluding items or persons from an analysis: items can be excluded by number; groups of persons can be specified using the candidate codes. One large matrix is then built from the individual test score matrices.

Analysis proceeds according to options selected by the user (regarding stopping values, etc: see Fig. 4.6). After each analysis the tables of estimated person and item values can be sorted in various ways and inspected, allowing decisions to be made on which items or persons to reject; the analysis can then be repeated.

Linking of newly-calibrated items into the bank is done at the end of the Rasch analysis routine.

4.3.3 Other facilities

System users can control to a certain extent how the bank works for ordinary users (teachers).

They can set the scale used for reporting difficulty and ability, exclude particular item types, specify exactly what gets printed when teachers print a test, and so on.

They can also define new CAT tests (see below).

4.4 Printed output from ItemBanker

The appendix at the end of this chapter shows a sample of the printed output from Item Banker: a question paper, an answer key, and a transformation table allowing raw scores to be interpreted directly in terms of bands on the user-supplied scale.

The appendix also shows an additional useful piece of output: the Personal Diagnostic Chart (generally called a *kidmap* in the Rasch literature, e.g. Pollitt 1990:881). This shows the scale of possible raw scores, set next to their equivalents in terms of the user-supplied scale. Then on each side the numbers of the items in the test are located on the scale according to their

difficulty. One thing to do with the chart is to ask students to mark those items which they got wrong, even though these items were below their estimated proficiency level. Such items may represent particular problems that would repay individual study.

The test shown in Appendix 1 contains 30 items. It may be seen that the transformation table does not report band scores for extreme high or low raw scores, signalling them only as being *off-target*. Similarly, the personal diagnostic chart cuts off extreme raw scores. The reason for this is simply that band scores (that is, ability estimates) for persons who get extreme low or high scores are distorted away from the mean: high scorers are estimated too high, low scorers too low. This points to a significant problem with the Rasch model, at least in the case of the present study of language proficiency testing. The same problem emerges with the estimation of item difficulty from poorly-targetted items, which again have extreme raw scores; it will be taken up in the chapter on estimation, and again in the concluding discussion. The proportion of off-target scores to exclude from the transformation table was settled by intuition, supported by some empirical evidence (see Chapter 7).

4.5 The Computer-adaptive test

The computer-adaptive test (CAT) is a self-standing program which uses Item Banker's data files. It allows learners to test themselves by interacting directly with the computer, typing in answers at the keyboard. Depending on the responses, the computer adjusts upwards or downwards the difficulty of the questions it asks, to that level where the learner has a 50 per cent chance of answering correctly. In this way the maximum amount of information is gained from each question. The test normally terminates when the estimated error falls below a certain level, which in practice takes about 18 to 20 questions.

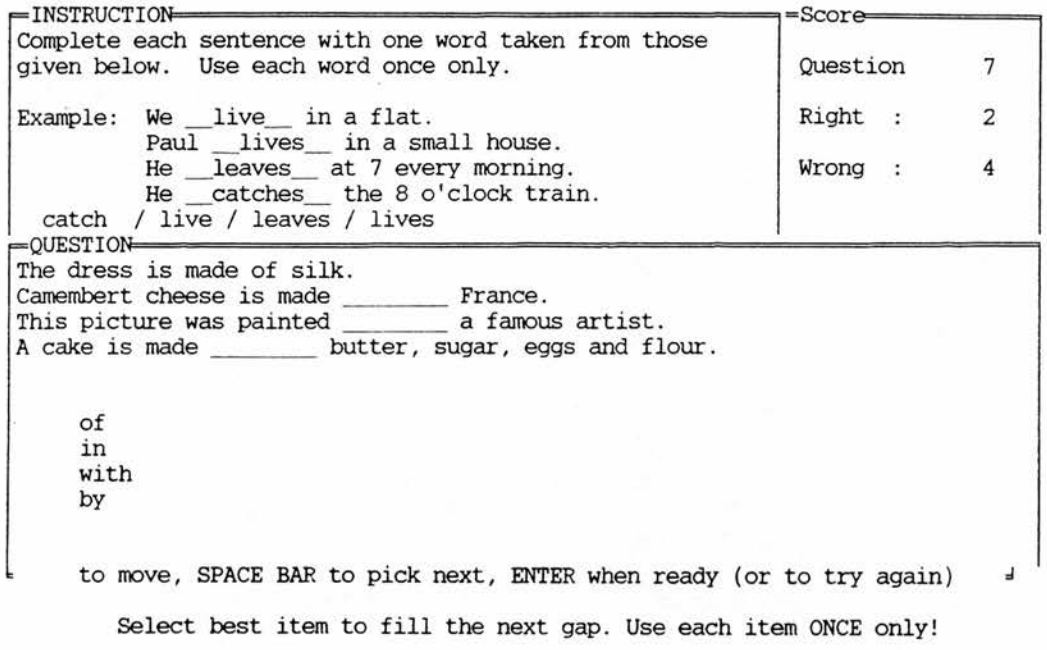


Fig. 4.7. Item Banker CAT: answering a matched pairs item

After entering their name, candidates are offered a choice of test. A CAT test is defined by a system user from within Item Banker, the definition comprising a selection of test content headings and item types. Level is not included in the definition, as of course the whole idea of computer-adaptive testing is that the level is found interactively in the course of the test. CAT test definitions are saved to a file, which is then used by the CAT program. All the items in the bank which match the definition are then available for use in the test.

The screen shows the instruction and the question prompt in a form resembling as closely as possible their appearance on paper. Some further instruction of specific relevance to the CAT may be given at the bottom of the screen (fig. 4.7).

The way of making a response should, ideally, be as similar as possible to the form of the paper test, given that item difficulties are estimated from trialling on paper, and these difficulties should remain the same (Wise 1989 reviews research on administering tests by CAT). Phrase or sentence types, or the one-word gap fill, seem to present identical problems on paper as on the computer (except, of course, for the need to type rather than write, which may disadvantage some and advantage others).

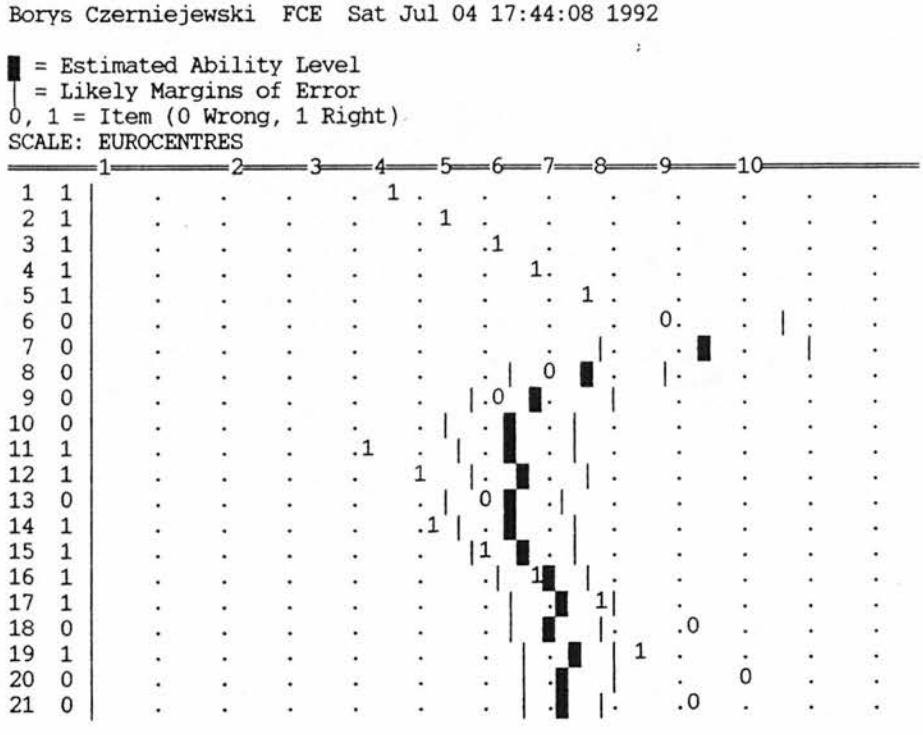


Fig. 4.8. Item Banker CAT: Example of printout showing progress through CAT test

The recognition types of items are more problematic. The solution of jumbled words and jumbled sentences types by writing numbers seems in any case rather cumbersome on paper, and would conceivably be even moreso on screen; thus it was decided to take advantage of the computer's possibilities, and allow learners to physically unjumble the elements. Whether this makes these items significantly easier is an empirical question which the present study will not be able to address.

After each response the program re-estimates the learner's ability, and the associated error using an algorithm given in Henning 1987:139.

When the test ends the learner's estimated level is reported in units of the user-supplied scale, and the program returns to the start. A disc file records output from the test, for possible later use. This includes:

the name of the candidate and the date and time of the test;
the text of each question, and the candidate's response;
the assessment (right or wrong).

Finally a chart showing the progress of the test is output (Fig. 4.8). It shows how the estimate of ability becomes increasingly stable, and the margins of error decrease.

4.6 Appendix: A sample Item Banker test

ItemBanker Test: Tenses & verb forms, Levels 5 - 9

Your Name: . . . . . .

Complete the sentence with a suitable phrase.

Example: A: What..... get up?

B: At 7 o'clock, usually.

You write: *time do you*

1 . .
Ian: Tom went to London yesterday.

Keith: Why ..... ?

Ian: On business, he said.

2 . .
A: I might be home late tonight.

B: Well if you ....., please don't make a lot of noise.

3 . .
Ian: I once ate raw fish.

Keith: Good heavens! Where ..... ?

Ian: In Japan.

4 . .
Policeman: When do you think your purse was stolen?

Lady: In the market, while I ..... some vegetables.

5 . .
A: How long have you been studying English?

B: By next June I ..... English for three years.

Mrs X: How long has your husband worked in the bank?

Mrs Y: By next May he ..... there for thirty years.

Complete the sentence with the correct form of the verb in brackets.

Example: He ..... on a farm last year. (WORK)

You write: *worked*

7 . .
I may go to the disco if she ..... me. (INVITE)

8 . .
When George got back from the pub last night, he found that his wife ..... his dinner to the dog. (FEED)

9 . .
A: Are you having steak?

B: I think ..... the salmon instead. (TRY)

10 . .
His mother said he was not at school because he .... by a dog the previous day. (BITE)

11 . .
By the time you get to London, the sun .....
(SHINE)

12 . .
By next June, he ..... English for three years. (STUDY)

13 . .
She dropped her wedding ring, which ..... across the floor and under the fridge. (SLIDE)

14 . .
Tom was playing volleyball on the beach when he _____ on some glass and cut his foot. (TREAD)

Write ONE word only to complete the sentence.

Example: Is _____ a bank near here?

You write: *there*

15 . .
Tim: You walked home? Why didn't you catch a bus?

Paul: Because I _____ to take any money.

16 . .
_____ you be coming to the office party tonight?

17 . .
No _____ had we dropped the match than the house burst into flames.

Make a question using the word in brackets.

Example: They are here. (WHY)

You write: *Why are they here?*

18 . .
He brought his cat. (WHY)

19 . .
It's been raining. (HOW LONG)

20 . .
They do their work. (WHERE)

21 . .
They are at home. (WHEN)

22 . .
He'd like to visit Spain. (WHEN)

23 . .
She lost it. (WHERE)

Finish the second sentence so it means exactly the same as the first one.

Example: When is she coming?

Do you know ..... ?

You write: *when she is coming*

24 . .
Send a telegram first thing on getting your results.

Send a telegram as soon .....

25 . .
"I went to the disco last night," said Ben.

Ben said that ..... .

26 . .
We will have collected the luggage by 9 pm.

By 9 pm the luggage .....

27 . .
I was late because I overslept.

If ..... late.

28 . .
The policeman asked me where I did my shopping.

"Where ..... shopping?" asked the policeman.

29 . .
He realised that someone had broken into his car.

He realised that his car..... into.

30 John didn't look after his car and now it has broken down.

If John .....

ItemBanker Test: Tenses & verb forms, Levels 5 - 9

ANSWER KEY

```
1 / 71
 did  he  go *
 was  that *
         do  that *


2 / 189
 are *
 JMP  come  home  late *
 do ..


3 / 72
 did  you  eat  it *
 was  that *
          do ..


4 / 182
 was  getting *
      buying *
      looking  for *
               at *


5 / 253
 shall  have  studied *
 will ..
 'll ..
              learned *
              been  studying *
                    learning *


6 / 1002
 will  have  worked *
             been  employed *
                   working *
                   JMP *


7 / 770
 invites *


8 / 964
 had  fed *
 was  feeding *
```

```
9 / 898
 We'll  try *
 I'll ..
 I  will ..
 I'm  going  to ..
    am ..


10 / 760
 had  been  bitten *


11 / 965
 should  be  shining *
 might ..
 will ..


12 / 248
 would  have  studied *
 will ..
              been  studying *


13 / 956
 slid *


14 / 957
 trod *


15 / 983
 forgot *


16 / 897
 Might *
 would *
 will *


17 / 908
 sooner *


18 / 165
 Why  did  he  bring  it *
                      his  cat *


19 / 171
 How  long  has  it  been  raining *


20 / 29
 Where  do  they  work *
                 do  their *
```

```
21 / 27
 When  are  they *


22 / 61
 When  would  he  like  to  go *
                          visit  Spain *
                                 JMP *


23 / 167
 Where  did  she  lose  it *


24 / 454
 as  you  have  your  results *
     you've  got ..
     you've ..
          get ..
          can  after  getting ..
          have ..
               on ..


25 / 675
 he  had  been  to  the  disco  yesterday  evening *
 he'd  gone ..
                                 last  night *
                                 the  previous ..


26 / 750
 will  have  been  collected *
       be ..


27 / 746
 I  had  not  overslept  I  would  not  be *
    hadn't ..
                          wouldn't  have  been *


28 / 1001
 did  you  go *
 do ..
           do  your *
               the *


29 / 738
 had  been  broken *


30 / 747
 had  looked  after  his  car  it  would  not  have  broken  down *
                                  wouldn't ..
```

ItemBanker Test: Tenses & verb forms, Levels 5 - 9

SCORE TRANSFORMATION TABLE Scale: EUROCENTRES

Score	Ability
30	OFF-TARGET
29	OFF-TARGET
28	OFF-TARGET
27	OFF-TARGET
26	OFF-TARGET
25	8.5
24	8.5
23	8.0
22	7.5
21	7.5
20	7.0
19	6.5
18	6.5
17	6.0
16	6.0
15	5.5
14	5.5
13	5.0
12	5.0
11	5.0
10	4.5
9	4.5
8	4.0
7	4.0
6	3.5
5	OFF-TARGET
4	OFF-TARGET
3	OFF-TARGET
2	OFF-TARGET
1	OFF-TARGET

OFF-TARGET means test was too easy or difficult for learner, who should take a different test. It does NOT mean learner's level necessarily lies outside the scale shown.

ItemBanker Test: Tenses & verb forms, Levels 5 - 9

PERSONAL DIAGNOSTIC CHART Item numbers RIGHT	Score	Scale: EUROCENTRES Level	Item Numbers WRONG
		9.0	
	25		
14			14
	24	8.5	
13			13
12	23		12
11		8.0	11
6 17	22		6 17
	21	7.5	
10	20		10
		7.0	
	19		
	18		
		6.5	
	17		
5			5
30	16		30
29		6.0	29
28	15		28
4 23 27	14		4 23 27
26		5.5	26
9	13		9
8 22			8 22
2 3 21 25	12		2 3 21 25
1 7 15 16 24			1 7 15 16 24
18 19 20	11	5.0	18 19 20
	10		
	9	4.5	
	8		
	7		
		4.0	
		3.5	
	6		

Draw a line across the page through your SCORE.
Read off your estimated LEVEL.
If your score isn't shown, try a harder or easier test.
Circle numbers of any items ABOVE your level which you got RIGHT.
These are your unexpectedly strong points!
Circle any items BELOW your level which you got WRONG.
These may be problems you need to work on.

5: Data collection, item calibration & model fit

This chapter follows the process of collecting data to construct the language proficiency variable. It begins with the writing of items, their inclusion in test forms, the trialling of these tests upon learners, and the checking and correction of this data. It then describes the approach taken to Rasch-analysing the data, and discusses practical problems encountered at this stage. Approaches are discussed to testing the quality of item calibration. Finally we examine the extent to which the calibrated items can be held to 'fit the model' - that is, to delineate a single, uniform language proficiency trait.

5.1 Data collection

5.1.1 Writing items

The first 500 items were assembled by the present writer. The starting point for item construction was a number of documents used in Eurocentres schools as guidelines for designing course content at each proficiency level. These guidelines included lists of language functions and grammatical structures. Some use was made of sets of existing test items developed earlier for use at various levels. The proficiency range aimed at was from beginner level to very advanced.

The selection of items was as inclusive as possible, taking in the following components of language competence (in terms of Bachman's (1990:87) taxonomy) :

> all components of grammatical competence except phonology (vocabulary, morphology, syntax, graphology);

all components of textual competence (cohesion, rhetorical organisation);

certain components of illocutionary competence (ideational functions, manipulative functions);

certain features of sociolinguistic competence (sensitivity to register, to naturalness).

The rationale for this decision has already been discussed (2.3.2 above). As a basis for trait construction we adopt the weak view of General Language Proficiency as an aggregate sort of measure based on as wide a range of evidence as possible. That is, we take the view that 'we must give testees a fair chance by giving them a *variety* of language tests' (Alderson 1981b:190). Of course, the amount of variety possible within a discrete-item, paper-and-pencil testing format is severely limited in respect of *method*, and hence of the language skills tested, yet in terms of content we see no reason not to be inclusive. We hypothesize on the grounds of the strength of the general factor in different measures of language proficiency that a heterogeneous collection of items will fit satisfactorily to a single dimension (a *measurement* dimension, to recall the distinction made above (3.3.2) in the discussion of unidimensionality).

Of course, we expect that there are limits to what will fit to the overall language proficiency trait. One purpose of including a wide variety of items is to discover what those limits are - the sort of testing to destruction suggested by Spolsky in his review of the ITESL:

The authors use their results to argue for the existence of a grammatical proficiency dimension, but some of the items are somewhat more general. The nouns, verbs and adjectives items for instance are more usually classified as

> vocabulary. One would have like to see different kinds of items added until the procedure showed that the limit of the unidimensionality criterion had now been reached. (Spolsky 1988:123)

We do not take the view that a trait so broadly conceived is necessarily uninterpretable. While the approach described may be seen as neglecting a necessary stage of *a priori* construct validation (Weir 1988), we believe that the investigation of item difficulty in Chapter 6 shows that *a posteriori* explanation is possible.

Most importantly, explanation is properly applied to *subsets* of items, rather than the bank as a whole. The trait we attempt to construct by including a heterogeneous collection of items can best be seen as a kind of matrix in which a number of theoretically interpretable traits may be fixed. They become interpretable precisely because they are fixed in a wider context. An example already introduced (2.2.4) is the *Accessibility Hierarchy* for relativization (Keenan & Comrie 1977). This proposed typological linguistic universal predicts an order of difficulty for relativizing different types of noun phrase. A set of items was constructed such that each type of relativization was included in two different item types, one simpler (providing the relativizing pronoun) and one more complex (a sentence completion task). It was predicted that when the item difficulties were found, the proposed order of difficulty would be replicated twice - in the simple items and again in the harder items. The results are discussed below (6.4.2). For now it is necessary simply to note that the results are all the more interpretable because the data points of the relativization problem are fixed on a single language proficiency scale: the distance between each point makes sense as an indication of the difference in difficulty. It turns out that the range of difficulty is very wide, so that to collect relevant data it is necessary to compare learners of widely-differing proficiency - a practical problem which is solved in the present case by using

IRT to construct the proficiency trait. The general proficiency scale provides an *interval* scale against which to study the relativization problem: without it one would be restricted to making *ordinal* comparisons.

At the stage of constructing the first 500 items it was hoped to include a number of sets such as this relativization one, in which particular language problems would be systematically investigated through manipulation of item features. This proposal proved difficult to carry through, however, given the accomodation which had to be made to the practical testing purposes served by the item bank. It is to these we now turn.

The item bank is seen as a practical testing and teaching resource. The 1000 items trialled at the stage of development reported in this study are not the end point: it is intended that the bank should continue to grow, to perhaps 2000 items. Nonetheless, the first 1000 items are expected to offer coverage of a wide range both of proficiency level and language content. Thus no one language area can be treated in much detail.

Two somewhat conflicting aims were identified, and it was attempted to satisfy both. On the one hand, it seemed that the bank should contain items with a clearly identifiable pedagogic point, so that learners' performance in tests should provide concrete feedback in the form of recommendations for remedial work, and so that teachers would have the possibility of using the bank to generate practice materials on particular areas of grammar, at appropriate levels (the risks of exploiting the bank both for teaching and testing are recognized, and will be discussed in the concluding chapter). For the reasons discussed above, it was felt that to concentrate exclusively on such items would tend to undermine the bank's value as a general proficiency measure. Hence the need for a second, less pedagogically-transparent sort of item: for example, concerning common idioms, collocations and lexically-determined syntactic problems. It was

decided that at the lower levels there should be more of the first type of item, with progressively more of the second type at higher levels.

After the first 500 items were written and in the process of trialling, the coverage achieved was examined, and the second 500 items were commissioned from a number of teachers in the Eurocentres group with the aim of filling identified gaps.

5.1.2 Linking of items in test forms

Items were included in test forms: 23 for the first 500 items, with about 27 items in each form, and 20 for the second 500, with about 32 items in each form.

Test forms were constructed so as to provide common item linkage in a basic block-diagonal pattern: that is, each test included a number of items from tests adjacent in level on either side. In addition there were four or five broad-range tests in each series, providing a second element of linkage.

Tests were trialled in schools of the Eurocentres group at various locations in England. Marking keys were provided, and the intention was that all papers should be returned ready-marked for analysis. Guidelines were provided for test administration, to minimize problems of data contamination through collaboration, time pressure, poorly-motivated performance or misunderstanding of the instructions.

5.1.3 Feedback during trialling, and revisions made

A number of problems emerged during trialling, which might have been mitigated if adequate pre-trialling of items had been undertaken. Time pressure unfortunately made this impossible.

There were a number of rather indifferent items. A particular problem was the large number of unpredicted responses to certain items. The answer keys were inadequate, particularly in the case of such unpredicted responses. Some test forms included too many different types of item. Particular question formats were criticized for being unclear, and these had to be modified. The layout and presentation of the first series of test forms was also generally thought to be unattractive.

The test forms for the first 500 items were produced by ordinary word-processing. By the time the second 500 items were assembled, the Item Banker software was in the process of development, and the second series of tests were produced directly from the bank, with an improvement in overall appearance (although the first generation of machine-generated answer keys was widely held to be unreadable).

Teachers were asked to provide detailed comments on problem items, and these were very helpful in making revisions.

5.1.4 Checking marking

The trials produced about 100 responses to each test form (minimum 85, maximum 130). In view of the problems with marking it was necessary to check all of the papers and ensure that a uniform marking scheme was applied. This was not only a time-consuming task, but also rather disspiriting, as it involved making a large number of arbitrary decisions as to which borderline cases were to count as right. However, the re-marking of papers led to some quite interesting observations concerning model fit, which will be discussed below.

Inspection of papers aroused suspicions of a certain amount of cheating, and a small number of papers were discarded at this stage; but the extent to which collaboration has contaminated the data remains of course unknowable.

Certain evidently flawed items were scrapped at the marking stage. Two types of item - the jumbled sentences, and the matched pairs type - were re-designed after the trialling of the first 500 items, the original format having been generally condemned as confusing. Close inspection of students' responses suggested that, confusing or not, there were very few evident problems, and it was decided to use the data from these items in the analysis, although it must be admitted that this is a possible source of distortion in the estimation of item difficulties.

In general it is impossible to estimate, in the present study, what influence on test performance such factors as the appearance and layout of the test papers may have had. The estimation of item difficulties has to proceed on the assumption that the difficulty of each item is unrelated to such things as its position in the test form, the layout, or the particular set of items in the form.

5.1.5 Retrialling

A number of items were quite badly targetted - that is, they were included in test forms for trialling with learners of inappropriate level. Such items were anwered nearly completely right, or wrong, and were thus badly estimated (see 5.2.3 below). Several new test forms were assembled for retrialling such items.

5.2 Item calibration

Calibration of items - that is, assigning each item a difficulty rating - is the first step in Rasch analysis of data. Where items are included in a number of different test forms for trialling, calibration involves adjusting the results found for each test form so that all items can be brought onto a single difficulty scale. Where the test forms cover different levels of ability/difficulty, this process can be called *vertical equating.*

5.2.1 Approaches to estimation

There are two basic approaches to item calibration involving the equating of several test forms.

The first is *common item equating.* Here each test form is Rasch-analysed separately, and difficulty ratings for items are found. Test forms for equating must contain a set of *anchor* or *link* items in common. The difference between the mean difficulty of these items found in the separate analyses constitutes the amount - the *translation constant* - by which one set of results must be adjusted to bring them onto the same scale as the other. Typically the difficulties of the first set of items analysed for inclusion in a bank are taken as fixed values, and subsequent items are adjusted to fit in with them.

The second approach is frequently called *one-step item banking.* It involves analysing all the test forms simultaneously. As a result of the analysis all items are placed directly on a single difficulty scale. Software for performing one-step Rasch analysis must be capable of dealing with a *missing data matrix* - a set of test scores in which each item is responded to by only a proportion of persons (and each person responds to only a proportion of items). In fact the algorithm is very similar to the simple maximum likelihood estimation algorithm, with the

difference that empty cells have to be excluded from the estimation procedure. The procedure depends, as in common item equating, on the test forms having items in common.

An advantage claimed for one-step item banking, apart from its administrative simplicity, is that the ability and difficulty estimates it provides are better - that is, that they reflect optimally closely the true state of affairs underlying the whole set of test scores. This is so because each item is calibrated in the context of responses from a wider range of persons, and each person is assessed in the context of a wider range of items and persons. Thus the maximum likelihood procedure should tend to estimates that delineate a trait - a continuum of difficulties and abilities - more clearly. But it is also the case that a missing data matrix is a more *unlikely* structure than the simple matrix, inasmuch as two elements with equivalent raw scores may not have equivalent difficulties/abilities. It therefore presents more work for the maximum-likelihood estimation procedure, with many more iterations necessary before stable values emerge, and a question mark possibly remaining as to whether the final set of estimates are sufficiently spread out from the mean (Lee 1991). There is conflicting evidence in the case of the present data and estimation algorithm. Some comparative tests were carried out on the performance of the one-step and common-item procedures; the interested reader is referred to the appendix at the end of this chapter, where this work is discussed.

A combination of these two approaches is of course possible: test forms can be analysed in groups, and each group linked in through common item equating to those already in the bank. This was the approach used in the present study.

5.2.2 Spread of difficulties

The first analyses of the first 500 items produced difficulty estimates covering a range with extreme values of -6 and +6 logits, although these values were evidently distorted. When poorly-estimated items are excluded from calibration, the basic bank profile is as follows:

No. of calibrated items	917
Mean item difficulty	-0.12
SD of item difficulty	1.45

Thus about 50% of items fall in the range -1 to +1, and about 90% fall in the range -2.4 to +2.4. The effective range of the bank is thus about 6 logits.

The original intention in writing items for the bank was to represent all levels of difficulty reasonably evenly, albeit with rather more items in the middle range. As the distribution of item difficulties is in fact almost exactly normal, we may suspect that the estimation procedure has failed to push the easier and more difficult items sufficiently far away from the mean.

5.2.3 Poorly estimated items

Extreme difficulty values were found for items with extreme raw scores (i.e. nearly all right or nearly all wrong), resulting from poor targetting. About 80 items in the first 500 seemed to be poorly estimated for this reason. The size of the distortion is shown, for example, in the fact that the most difficult items in the easiest, beginner-level test form were given difficulty estimates putting them at about upper-intermediate level. In effect, the difficulty scale found in any single Rasch analysis appears to be stretched at the extreme ends, losing the linear quality which is claimed to be one of the model's great

advantages. This would not be of much practical consequence in the case of a single, complete matrix of scores; but where missing data is involved, and vertical equating using common items is to be undertaken, it threatens to introduce considerable inaccuracy.

Subsequent analyses of the first and second 500 items excluded badly-targetted items at the outset. Excluding badly-estimated items produces different, presumably better, estimates for the remaining, good items. Fig. 5.1 illustrates this with data from a single test form: 23 items estimated both with and without the company of 5 bad items show slightly different values.

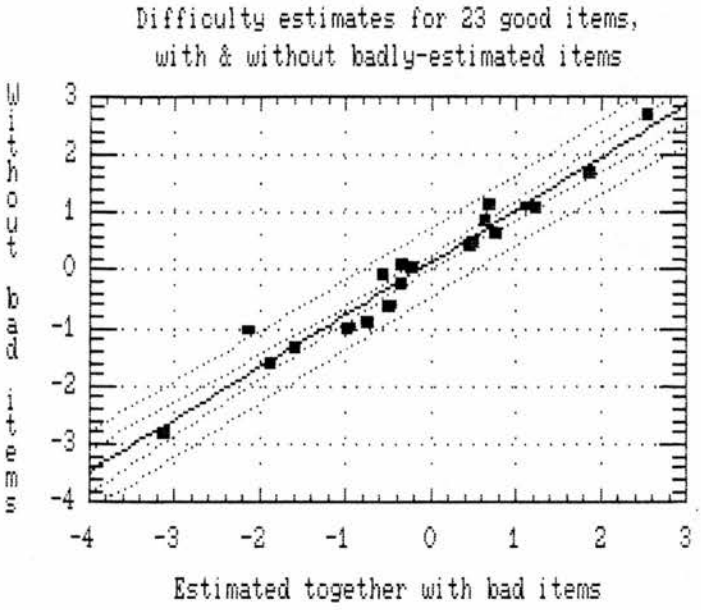


Figure 5.1 Two estimates of item difficulty: 23 items with and without the company of 5 badly-estimated items

Several test forms were prepared containing badly-targetted items. These were re-trialled. Fig. 5.2 shows comparisons of original and revised item difficulties for three of these test forms.

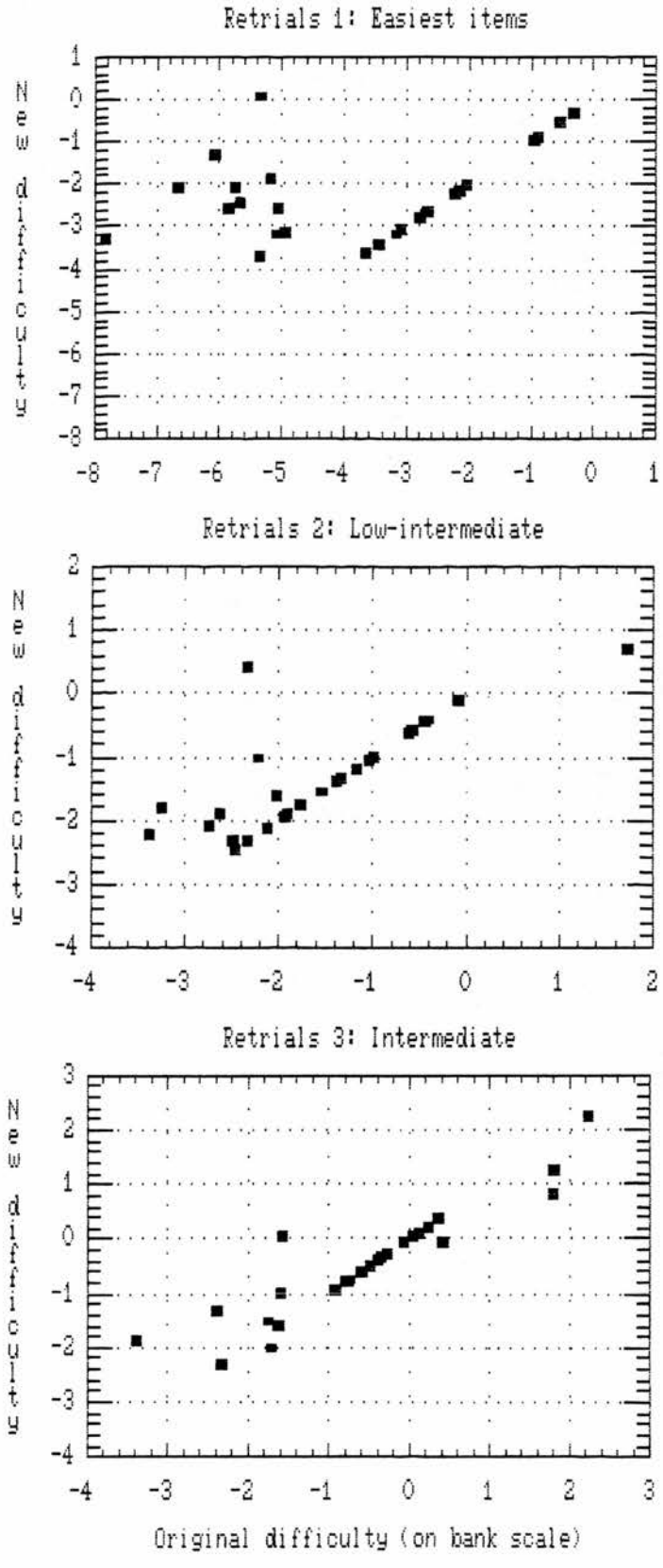


Figure 5.2 Comparisons of item difficulty: original estimates, and from retrials on learners of more appropriate level.

Each form contains items originally trialled in a number of different forms. The difficulties shown are the items' final location on the single bank scale after linking into the bank. The items lying on the identity line in each plot are anchor items included in the retrials to provide the link of common items into the bank: their difficulty is held constant.

The effect is most clearly visible in the plot for the easiest items, as the movement is all in one direction: all the recalibrated items were originally found too easy, and their original estimates were thus biassed downwards. When retrialled they take on values nearer the centre of the scale. Plot 2, for the lower-intermediate items, shows the same general effect, with one item too difficult for the original tests now adjusted downwards. Plot 3, at intermediate level, is more difficult to interpret, because it includes items originally biassed both upwards and downwards, so we see movement in both directions.

These comparisons are complicated by sampling error. None the less, the general effect is evident enough, the difference between the two calibrations being in some cases dramatic. There is cause for concern here for anyone interested in using the Rasch model for vertical equating over a large range of ability. It does not appear possible in the present study to pursue the problem further, and yet it is certainly a problem which deserves investigation: particularly, we should like to know just how extreme the percentage-correct score has to be for the effect to become noticeable. In the present study it seemed that scores below 20% or above 80% deserved treating with suspicion.

It seems to be primarily the removal of extreme scores which accounts for the relatively short scale length achieved. 6 logits, the effective range of the bank items, is less than is generally reported for scales covering a wide ability range. By

setting limits on the data we attempt to fit, the Rasch model can be made to perform satisfactorily, but at a cost in terms of the scale length.

5.3 Investigating model-data fit

Hambleton and Swaminathan (1985) discuss three general ways in which to determine "how well a model accounts for a set of test data".

> 1. Determine if the test data satisfy the assumptions of the test model of interest.
>
> 2. Determine if the expected advantages derived from the use of the item response model (for example, invariant item and ability estimates) are obtained.
>
> 3. Determine the closeness of the fit between predictions and observable outcomes (for example, test score distributions) utilizing model parameter estimates and the test data.
> (Hambleton & Swaminathan 1985:151)

The following discussion will follow this general outline.

5.3.1 Testing model assumptions

Hambleton & Swaminathan discuss four assumptions made in applying the Rasch model, concerning unidimensionality, equal item discrimination, the absence of guessing, and the absence of speededness.

Unidimensionality

As Hambleton & Swaminathan point out, the major assumption made in applying an item response model is that of unidimensionality. They describe approaches to checking on the unidimensionality of the data, using factor analysis among other things. But it is questionable whether the proposed techniques achieve anything which Rasch misfit analysis itself fails to. Smith (1991) compares factor analysis and Rasch misfit analysis in terms of their ability to detect violations of unidimensionality, and concludes that for most practical purposes misfit analysis is of more use. Thus we consider that the unidimensionality question is adequately treated under the discussion of misfit (below, 5.3.3).

Equal discrimination

A second assumption made by the Rasch model (in contrast to the two-and three- parameter item response models) is that items should discriminate equally. Hambleton & Swaminathan find only descriptive methods available for assessing the viability of this assumption, and propose comparing the range of item point-biserial correlations as a 'rough check' (p.159). Test form 22, chosen at random for analysis, shows point-biserial correlations ranging from .01 to .61, with an average of .35 and a standard deviation of .13. In the absence of a basis for comparison, one cannot really say whether this is 'small' or not.

Assumptions of no guessing, no speededness

The no guessing assumption can be reasonably safely made in the case of the present study, because selected-response item types make up only part of the items, and even here, the random chance of successful guessing is small (1 in 24). This is because the classical multiple-choice four or five choice format is not used.

The assumption that test administration was not speeded was not checked by any of the methods proposed by Hambleton & Swaminathan. However, teacher reports indicated that time pressure was not generally a problem, and in those few cases where it was evident, (where students failed to finish the paper) scores could be treated as missing, rather than wrong.

5.3.2 Checking model features (Invariance of estimates)

A major advantage claimed for item response models is that estimates of item difficulty are 'sample free' - that they do not depend on the particular group of persons used in trialling. Several approaches were tried to see how far this property obtained in the present case.

Item difficulties estimated from split groups

Figure 5.3 shows four comparisons of item difficulty estimates obtained from different groups of persons, for Test 33 (selected at random for analysis). The groups compared, and the results obtained, are summarised below:

	X-axis	Y-axis	r (corrected)	slope
a)	All persons	Top 50%	.94 (.79)	1.03
b)	All persons	Middle 50%	.98 (.95)	1.006
c)	All persons	Bottom 50%	.97 (.83)	1.02
d)	Bottom 50%	Top 50%	.86	0.89

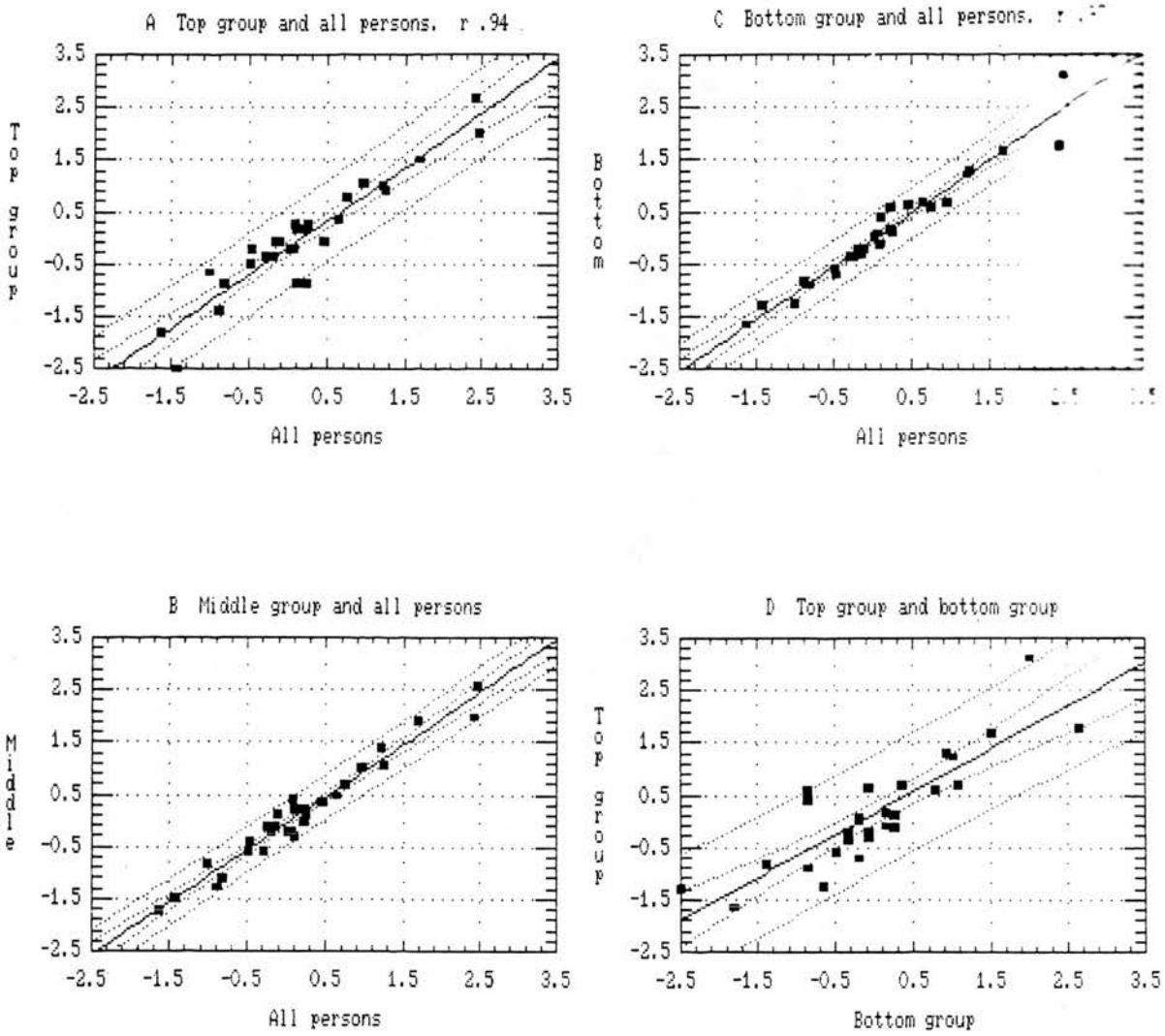


Figure 5.3 Comparisons of item difficulty estimates obtained from high, mid and low-level groups of persons

One problem for this comparison, as for others that involve splitting candidates into groups, is that the total sample size for each test paper (about 100 persons) is too small for analyses based on subsets to be very accurate. This shortcoming was unavoidable, given that the present study concentrated on obtaining calibrations for a rather large number of items, with a limited population of learners available for trialling. Some of the spread visible in these comparisons is thus due to increased error from the smaller samples.

The first three cases compare part of the population with the whole, and therefore manifest a significant amount of part-whole correlation. The figure shown in brackets applies a correction for part-whole overlap, calculated from person raw score totals.

As would be expected, the agreement is closest for case b), the whole group compared with the middle 50%. The interesting case is d), the top group compared with the bottom. According to the corrected figures, the agreement appears to be about as good as the others, a result which is quite encouraging. The slope of the regression is, however, further away from the value of one which it should have if the scale length in the two estimations were the same. From the above discussion of extreme scores and the threat they represent for estimation, we can see that this flattening of the slope is consistent with a degree of scale stretching: for the top group it is the easy items which show more extreme values, while for the bottom group it is the harder ones.

Item difficulties independently estimated

Figure 5.4 illustrates a different check on the invariability of item estimates. Four sets of items were identified which had been included in more than one test form, in such a way that estimations could be found from entirely independent test administrations and Rasch anâlyses. Thus, for example, at the lowest level 8 items were found in the group of tests 1 to 6 and

20 which could also be estimated from tests not in this group. The extent of agreement again appears reasonably encouraging, although again, the slope of the regression line is in most cases not unity.

Item difficulty and L1 background

Figure 5.5 shows another attempt to compare estimates of item difficulty derived from different splits of the person population, this time taking L1 background as the basis of the split.

Each pair of plots shows, on the left, a comparison of estimates from the whole group, and from a common L1 group identified within the whole group. On the right, for comparison, is a similar estimation done using a set, roughly the same size, of persons of other L1s, picked at random from the same group. This attempts to adapt a proposal by Angoff(1982), exemplified by Hambleton & Swaminathan (1985:180), for obtaining a 'baseline plot' to inform visual inspection: if the L1 group performed significantly differently, we would expect the L1 plot to show more dispersion than the comparable plot for the randomly-chosen set. Unfortunately, what appears to be revealed here is that estimation from small samples contains more error. Both the L1 and control groups show quite marked dispersion. What can be said is that no evident effect of L1 background emerges from this comparison. L1 background will be taken up again below, in the discussion of model fit.

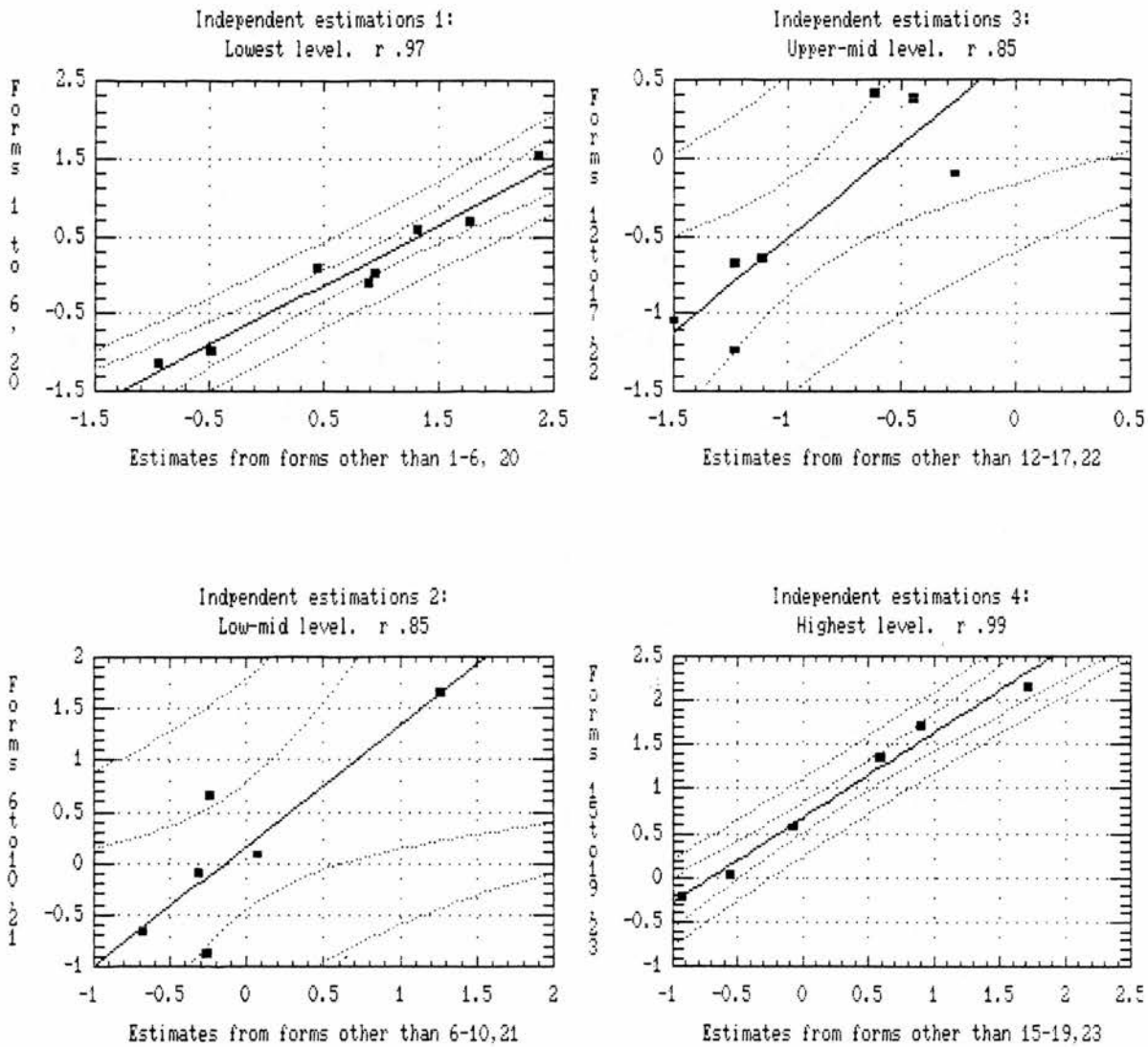


Figure 5.4 Four sets of items estimated independently

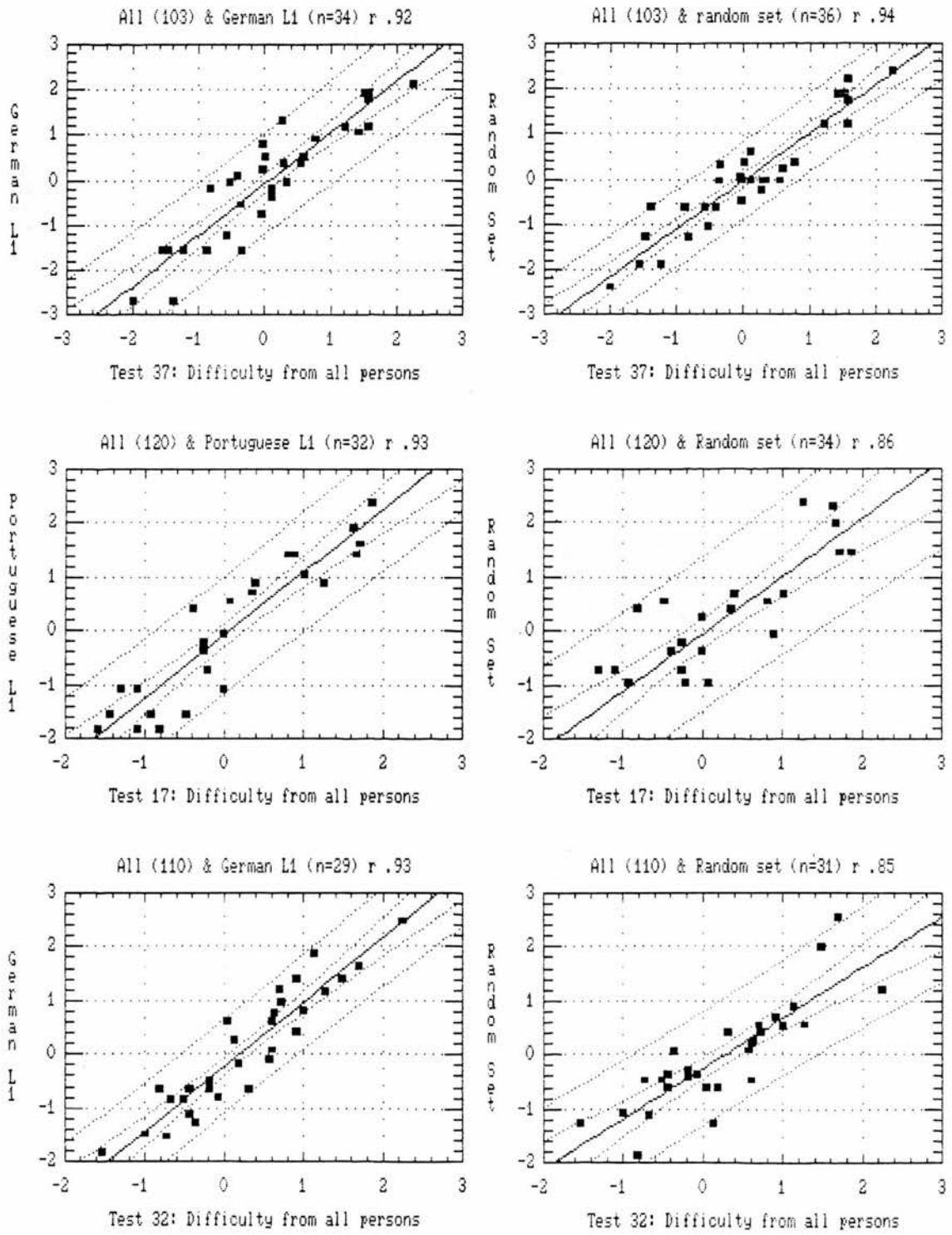


Figure 5.5 Item difficulties estimated from whole group, L1 group, randomly selected group

Actual and predicted item difficulties

One further investigation into the quality of item difficulty estimation concerned comparison of actual and predicted person performance.

Figure 5.6 shows, for two test forms, a comparison of actual and predicted outcomes at 11 different ability levels. Difficulty/ability is shown on the X-axis, while the Y-axis shows the proportion of items in the test answered correctly at each ability level. The dotted line shows predicted values, the heavy line those actually observed. Agreement seems good, although there is a small and consistent difference.

The same investigation was attempted for individual items, comparing proportions of people answering correctly at different ability levels. A problem here was again the relatively small number of persons, making it difficult to get enough cases in each ability category. An analysis with five ability categories was attempted on a number of test forms, but it was rare for cases to be sufficiently evenly distributed for any unambiguous interpretation to be possible. It had been hoped that this approach would enable the investigation of differential item discrimination, but with the given sample size this was not possible.

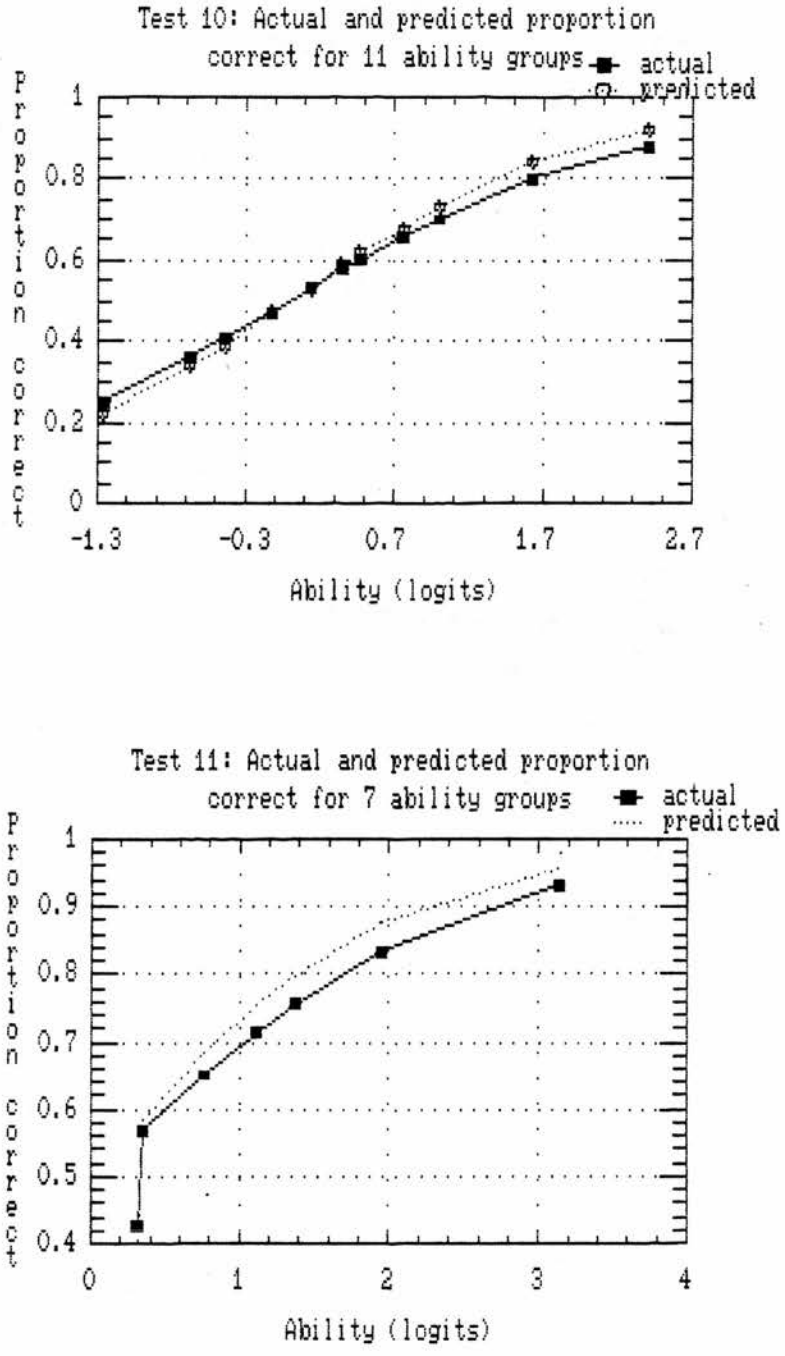


Figure 5.6 Actual and predicted performance on two tests

5.3.3 Investigating model fit

The investigation of model fit in Rasch estimation is where construct validition really starts. Items and persons are examined to see whether the observed pattern of responses is close to what would be expected, given the difficulty/ability rating found. That is, an item should regularly be responded to wrongly by less able persons, correctly by more able persons. If observed and expected responses differ by more than a certain amount (the limit is of course conventional) the item or person is said not to 'fit the model'. An examination of misfitting items may reveal what it is that they measure which is *different* from what the other items measure (and thus throw light on what it is that the other items *do* in fact measure). When the misfitting items have been removed, there is a sense in which the remaining items can be said to measure the 'same thing' - that is, they delineate a unidimensional trait.

Misfit, then, is manifested in deviations of observed outcomes from those predicted by the model. There are several ways in which such deviations can be measured. A standardized mean square statistic can be used to sum deviations for each item across all the people who respond to that item, or for each person across all the items he responds to. This statistic may be unweighted, in which case it is known as *outfit*, or weighted in order to lessen the effect of very unexpected responses (when for example an able person gets a very easy item wrong). The weighted statistic is known as *infit*.

The meansquare can also be expressed as a *t fit* statistic (Wright & Stone 1979, Henning 1987:123). The advantage of t fit is that it is a signed statistic which can indicate both lack of fit to the model (high positive values) as well as overfit (high negative values). Item Banker uses the unweighted outfit, and reports using *t*.

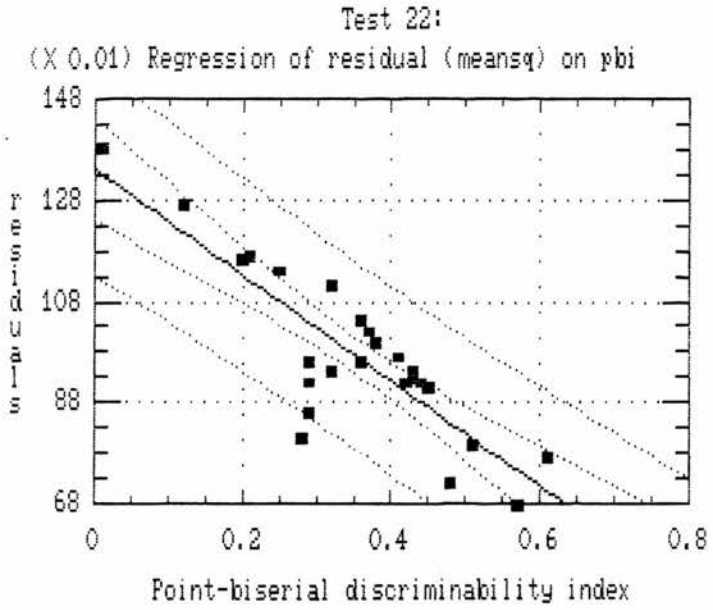


Figure 5.7 Comparison of point-biserial discriminability indices and Rasch fit (mean square) statistic

The notion of misfit in Item Response Theory evidently bears some relationship to discriminability in classical test theory, using such measures as point-biserial correlation. Figure 5.7 illustrates that while such discriminability measures do indeed relate strongly to Rasch model measures of fit, the correspondence is not perfect.

> With TTT (traditional test theory), it is generally assumed that the greater the discrimination the better, because this increases the reliability, even though it is known from the attenuation paradox literature of TTT (Loevinger, 1954), that a reliability can be too high - that is, increase in reliability beyond a certain point leads to a decrease in validity. ... Thus the SLM (Simple Logistic Model) formalizes the tension between reliability and validity which is dealt with only informally in TTT: When an item or items discriminate too highly relative to the other items, then it or they begin contributing redundant information relative to the other items and begin to increase the reliability at the expense of validity.

Andrich (1988:85)

In the present study very few items were rejected because of misfit, partly because the limits for what was considered acceptable were set quite high. This was in line with the practical orientation of the study - to produce a large bank of useable items. It also recognized the fact that given the small person sample size, and the less-than-perfect conditions in which some of the tests were administered, much apparent misfit might be due to factors unconnected with the quality of the items. It is also the case that misfit should be *interpretable*, if the rejection of misfitting items and retention of those that fit is not to become a simple capitalization on chance. While reasons for some cases of misfit can be found, it seems that most often the reason for misfit is not at all clear (which is not to say that there might not be a reason).

Generally speaking, it appears that language test items of the kind included in the present study fit quite readily to a single trait.

5.3.3.1 Alternative marking schemes and fit

If *absolute* fit of items was generally satisfactory, some experiments with alternative marking schemes enabled the comparison of *relative* fit of the same items marked according to different criteria, producing some suggestive findings.

As explained above, the purpose of examining the marking schemes was to make sure that all papers had been marked according to the same criteria. This procedure produced uncorrected and corrected score matrices for 12 test forms. Analyses of these were compared, primarily to check that the corrected ones (containing, presumably, less random error) fitted better than the uncorrected. While this was more or less the case, it was

evident that certain items actually fitted much worse after re-scoring. For example, accepting the deviant spelling of 'earlier', 'easier' as 'earlyer', 'easyer' made items fit worse, while *dis*allowing deviant spellings with omitted double consonants 'stoped', 'bigest' etc made items fit better. Such findings might be relevant to a discussion of validity. The above spelling example suggests that accuracy in spelling is coherent with the trait as a whole, which would in turn suggest that the trait has more to do with the development of language ability in a formal instructional setting (which is a reasonable guess). But in general, extending the range of acceptable answers beyond the narrow, 'correct' (in terms of a prescriptive, pedagogical rule), produces better fit. This makes sense, inasmuch as the criterion for giving a mark becomes less arbitrary, and thus reflects learners' true abilities better (rather than their skill or luck at divining the intention of the item writer).

5.3.3.2 Item types and fit

As described above, first estimations of item difficulty included some poorly-estimated items. Later estimations excluded these. Early analyses of different subsets of calibrated items, drawn from the bank, suggested a clear relationship between item type and fit, with certain item types fitting better than others. In later analyses the relationship is less evident. It should be remembered that in the process of calibrating items, items with high *positive* misfit are routinely removed, and so never make it into the bank. Items with high *negative* misfit, otherwise called *overfit*, are not removed. It appears that high overfit tends to be a characteristic of poorly-estimated items, and thus that when these are removed from the data, the overall fit statistics of the items in the bank are improved.

Table 5.1 shows the five main item types, with, under each type, the different question types or rubrics supported by that type. The SD of the t fit statistic is taken as an indicator of the spread of fit across items in a group. The table shows, for example, that the question type 'Complete the sentence with a suitable phrase' fits rather well (SD of t fit 1.18), and the type 'Supply one word to complete the sentence' fits worst (SD of t fit 1.47).

These differences fall well short of significance at the 95% level (using a ratio of variances test), and so can be seen as no more than suggestive. If the one-word gap fill performs worse, on average, than other item types, one might be tempted to attribute this to a greater influence of L1 in this kind of task. Pollitt & Taylor (1991) investigate question level bias in FCE cloze questions, and demonstrate the powerful and prevalent influence of L1 transfer. The similarity of the discrete-item gap fill and the cloze test would suggest that the same factor is at work in both cases. Pollitt & Taylor argue that:

> cloze is essentially a *productive language task*, in which context, cotext, syntax and grammar combine to set constraints within which the production process must take place. At an intermediate level, where students meet FCE, it must make a considerable difference to their chances if the constraints operate in a way that is familiar from their L1, or if they are forced to work only from a very imperfect knowledge of L2.
> Pollitt & Taylor (1991:7)

ITEM TYPES	SD t	mean diff.	n
Jumbled Words	1.19	-1.06	55
Jumbled Sentences	1.36	-0.69	43
Matched choices	1.33	-0.5	90
Gap fill	1.46	-0.05	234
one word	1.47	-0.17	168
correct form	1.38	-0.27	66
Phrase or sentence	1.25	0.07	488
suitable phrase	1.18	-0.72	76
correct form verb	1.38	-0.43	106
make question	1.13	-0.86	38
add words	1.28	0.19	36
use the word	1.30	1.08	55
transform	1.19	0.58	169

Table 5.1 Item types and fit .

5.3.3.3 Difficulty and fit

Early analyses suggested that items fit worse at lower difficulty/proficiency levels; but again, this effect disappears in later analyses with cleaner data. Table 5.2 shows the results of an analysis conducted on the whole set of calibrated items, divided into three more or less equal-sized groups by difficulty. Again, SD of t fit is used to indicate the spread of fit for items in each group. Fit is roughly the same over the whole difficulty range, the slightly greater spread at the lower and higher end being insignificant. The present study is thus

not able to confirm what other researchers have found concerning the dimensionality of language proficiency at different levels (Sang *et al* 1986, Milanovic 1988).

	Difficulty (logits)		SD t fit	n
Lowest (easy items)	< -.75		1.37	296
Middle	> -.75	< .6	1.25	336
Highest (hard items)		> .6	1.38	278

Table 5.2 Item difficulty and fit

5.3.3.4 Item content and fit

If items in the bank are grouped in subsets according to their language content tags, the fit of these content subsets can be compared. A grouping by just the 4 superordinate tags is shown in table 5.3. This is evidently a very approximate breakdown, as many items belong in more than one of these categories, and the tagging of each item may not of course correspond closely to what it actually measures. None the less, the picture presented in table 5.3 is readily interpretable. It shows that Grammar items lie closest to the proficiency trait depicted by the bank items. Textual items, dealing essentially with relations above sentence level, come next. Vocabulary and idiom items are rather less coherent with the trait, and Functional items, testing, let us say, sociolinguistic competence, are the least coherent of all. This would be nice were it true, but again, these differences in fit lack statistical significance; they remain, therefore, at best suggestive.

ITEMS BY CONTENT TAGS	SD of t	Mean diff.	n
Functional / Notional	1.48	-0.19	104
Grammar	1.31	-0.25	702
Textual	1.36	-0.30	61
Vocabulary & idiom	1.39	0.6	223

Table 5.3 Item content and fit

5.3.3.5 The influence of L1 on fit

The present study also attempted to detect in misfit the influence of L1. As described above, a comparison of calibrations from groups split on the basis of L1 failed to demonstrate that L1 had any clear effect on the estimation of item difficulties.

A different approach followed Pollitt & Taylor (1991) and involved analysing the residual matrix to calculate estimates of item bias for different L1 subsets. A number of test forms were Rasch analysed individually, and for each form the larger L1 groups were identified (data were generated for groups of 13 persons or more). For each item a measure of bias was found, indicating to what extent the item was easier or more difficult for each L1 group than it was for the group as a whole. Standard errors were generated for each bias measure, allowing significant effects to be identified.

This approach makes possible an investigation of item bias for those L1 groups which happen to be reasonably well represented in any particular test administration. It is rather unsystematic, but it is the best that could be attempted with the available data.

11 test forms were analysed. Each form contained about 32 items. In the data for each form between 2 and 5 L1 groups of sufficient size were identified. This gave a total of 1157 group/item observations. Taking a 95% significance level we would expect that random chance alone would produce 58 spurious 'significant' observations. In fact, 65 observations were found to be significant. This is not a dramatic result, and it indicates that caution is necessary in attempting to interpret apparent cases of L1 bias. Particular examples reinforce this impression. One item found to be significantly more difficult for Portuguese speakers in one test administration was found *easier* for the same L1 group (though not significantly) when it appeared in another test form. Given the small size of some L1 groups, random effects, or collaboration among common-L1 colleagues in a single class of learners, might be enough to explain many observations.

A limited investigation of apparently biassed items was undertaken, looking at the more significant cases (statistically speaking). Many cases are in fact readily interpretable in terms of L1 bias, as the following examples may show.

1 The red blouse costs more _____ the white one.

This item appears 2.8 logits more difficult for Japanese L1 (n = 13). Nearly all Japanese L1 supplied the word 'expensive'. The syntax of expressions of comparison in Japanese is very different, whereas European languages have quite similar structures.

2 I / weekend / visit / at / parents / my / the / often.

3 Peter / to / train / goes / always / work / by.

These word-ordering items are 2.1 logits and 1.1 logits harder respectively for German L1 (n = 26). For item 2 a frequent answer was 'I visit often...'. For item 3 the most frequent wrong answer was 'Peter goes always to work by train.' This seems to reflect transfer of adverb positioning in German.

4 The man _____ you saw running away was the murderer.

This item was 2.8 logits harder for German L1 (n = 26). All Germans who answered wrongly supplied 'which'.

5 I _____ go by taxi. (RARE)

This item was 1.3 logits harder for German L1 (n = 30), although in another test form the same item was no harder for the German group (n = 13). This is a further caution against over-interpreting the data. Germans offered a wide range of wrong answers, but it is hard to say why.

6 He plays football. (HOW OFTEN)

7 It takes about ten minutes. (HOW LONG)

These question-formation items were 1.3 logits and 1.2 logits harder respectively for German L1 (n = 26). A variety of wrong answers include question-formation by inversion - 'How often plays he football? - but the most common error is to omit any syntactic marking of the interrogative: 'How often he plays football?' It appears that DO-question formation is more difficult for German speakers, but this does not manifest itself strongly through direct transfer of an L1 syntactic pattern. The same seems to be true of the following two verb-tense items.

8 When I arrived, Mary .... (ALREADY LEAVE)

9 A: What are you doing this afternoon?
B: I ..... tennis with Sam. (PLAY)

Both these items involve supplying the correct form of the verb, and were 1.3 and 1.1 logits harder for German L1. No single wrong answers seemed characteristic for Germans, however.

Bias is of course a relative notion, and some cases of items being *easier* for a given L1 are best interpretable as meaning that the items are *more difficult* for the other L1 groups. For example:

10 His hair ..... long, black and curly. (BE)

This item is 1.3 logits easier for Japanese L1 (n = 13) - possibly because Japanese speakers are *not* tempted to supply a plural form of the verb.

This brief investigation of possible L1-related bias as a factor in explaining item misfit suggests the tentative conclusion that such bias is present, but is a less prevalent phenomenon than found by Pollitt & Taylor (1991). With small samples in particular, spurious significance may be a problem. However, the utility of Rasch misfit analysis as an approach to investigating item bias is well demonstrated.

5.4 Appendix: a note on Rasch estimation of item difficulties

Two approaches to calibrating items trialled in a set of linked test forms have been mentioned: the *common-item* approach and the *one-step* approach. It was noted that the one-step approach is generally preferred, as it should tend to delineate a trait more clearly; at the same time, there remains the problem of knowing how many iterations are necessary for estimates to converge, and whether the final set of estimates are sufficiently spread out from the mean (Lee 1991).

To get a better picture of how estimation works, some tests were carried out on single test forms and linked sets of forms, using in the latter case both common-item and one-step methods.

In each of the cases discussed below we study the *difference* in item difficulties found, either between two estimations using a different number of iterations, or between two methods, to a similar level of precision. Precision here is understood as the *stopping value* specified, i.e. the smallness of the sum of adjustments (the closeness to convergence) that must be reached before the algorithm stops. For example, in this study the following values for precision were used: low (<.001), medium (<.0001) and high (<.00001).

Fig. 1 shows the simplest case: the estimation of difficulties for a single test form (test 28), i.e. from a matrix with no missing data. In this case, values converge quickly: 3 iterations reaches the lowest precision, and 4 reaches the highest. The difference between these iterations is shown in Fig. 1. The X-axis shows item difficulty in logits, the Y-axis the difference in difficulty, also in logits (note the scaling factor). We see that the scale length has increased: easier items are easier, harder items harder, hence the slope of the

line. The differences are however very small, running from -0.0006 logits to +0.0014 logits; so this estimation may be considered to have converged.

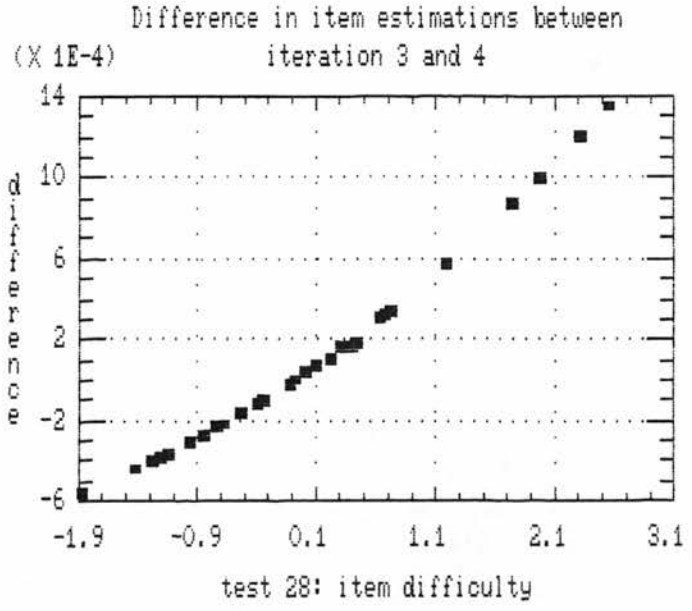


Fig. 1. Simple estimation of one test form (Test 28)

Fig. 2 shows the case of two test forms (test 28, 38) estimated by the one-step method, i.e. from a missing-data matrix. With missing data, more iterations are required for a given level of precision: in this case, 4 iterations reach medium precision, 13 reach high precision. The difference between these two estimations is shown in Fig.2. The two tests appear very clearly as straight, nearly horizontal lines. The lower line slopes very gently, but practically speaking *in one test, the difficulties of the items have not changed relative to each other.* In other words, the relative difficulties of the items in each test form have converged to stable values. Thus it is *the tests as a whole* which are being pushed further away from each other by successive iterations. The items doing the pushing are the 5 common anchor items, which can be seen in the middle of the plot.

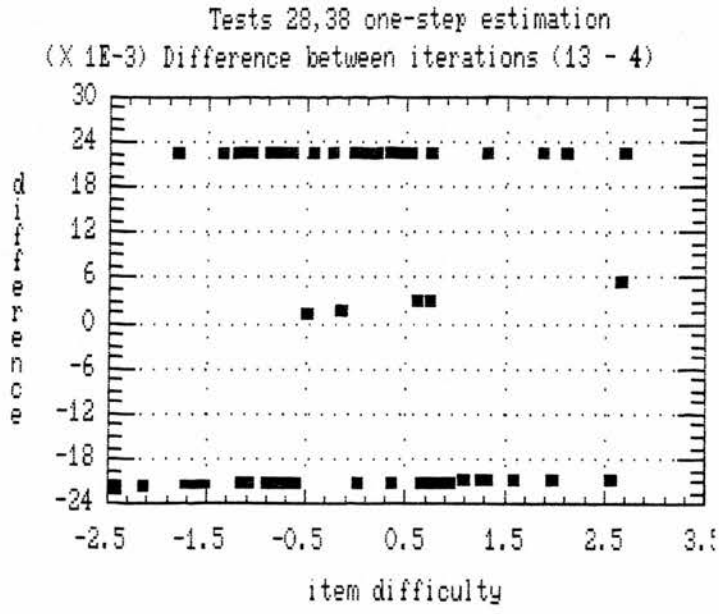


Fig. 2. Estimation of two forms (28, 38) by One-step method

One suspicion had been that with more iterations it was the poorly-estimated items in the tails of the distribution for each test form which were taking on increasingly extreme difficulty values. Fig. 2 seems to show that this is not the case: all the items in a test move in unison.

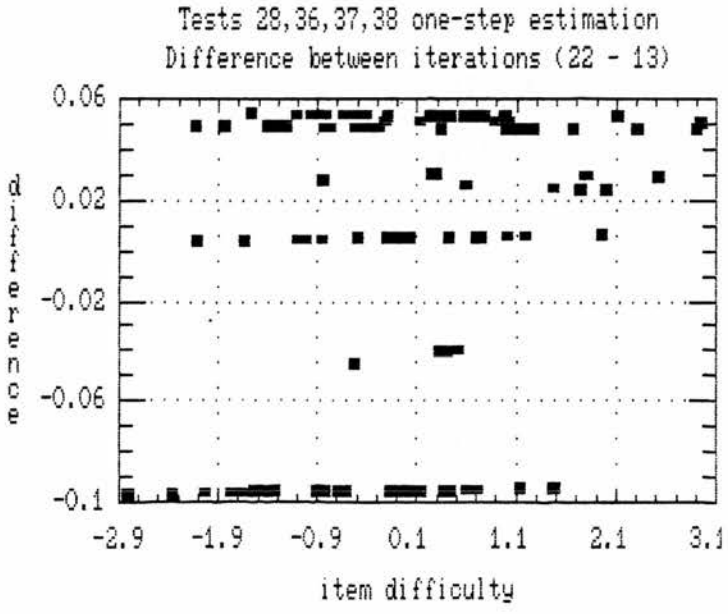


Fig. 3. Estimation of four forms (28,36,37,38), one-step method

Fig. 3 simply illustrates the same effect for a more complex missing-data matrix containing four tests: test 28, 36, 37, 38. Again, each test shows up as a straight line (two of the tests are close together). The anchor items are scattered between, exerting outwards pressure.

Again, the size of the difference is not great: in Fig. 2 it is at most .03 logits, in Fig. 3 at most -0.1 logits (it can be seen that the more complex the matrix, the greater the difference between the two levels of precision). Thus it seems unlikely that the scale would grow very much longer, even given a higher precision level and a much larger number of iterations. Even in the present study, where two large series of test forms were arranged with block-diagonal linking to cover the widest possible proficiency range, the series of increments in scale length that might be achieved seem unlikely to add up to a significant difference in total scale length.

One way to establish how much further the estimates from a missing-data estimation might stretch should be to compare them with the results of a common-item estimation. It seems reasonable to suppose that the *translation constant* - the mean difference in the difficulties found for the common anchor items in separate estimations of two different tests - represents the best possible estimate of the true difference in the difficulty of those tests. If a one-step estimation can be shown to separate two tests by as much as the translation constant in a common-item estimation, then we should be satisfied that the estimation has converged to stable values. A measure of this separation is easily achieved, by comparing the mean difficulty of each test in the one-step analysis. Doing this for tests 28 and 38 we find:

	Precision:	*< .0001*	*<.00001*
Mean difficulty:	test 38	-.0960	-.1172
	test 28	.0818	.1013
	difference	.1779	.2185

Thus the best estimate of separation is 0.2185. The translation constant from a common-item estimation is 0.2119. This would suggest that the one-step estimation has separated the tests satisfactorily.

However, the reality appears to be slightly more complex. Fig. 4 compares difficulty estimates for the pair of tests 28 and 38, but this time the difference is not in the number of iterations, but in the estimation method. Difficulties found by common-item equating (also meaned to 0) are subtracted from difficulties found by the one-step method. Precision was <.00001 in both cases.

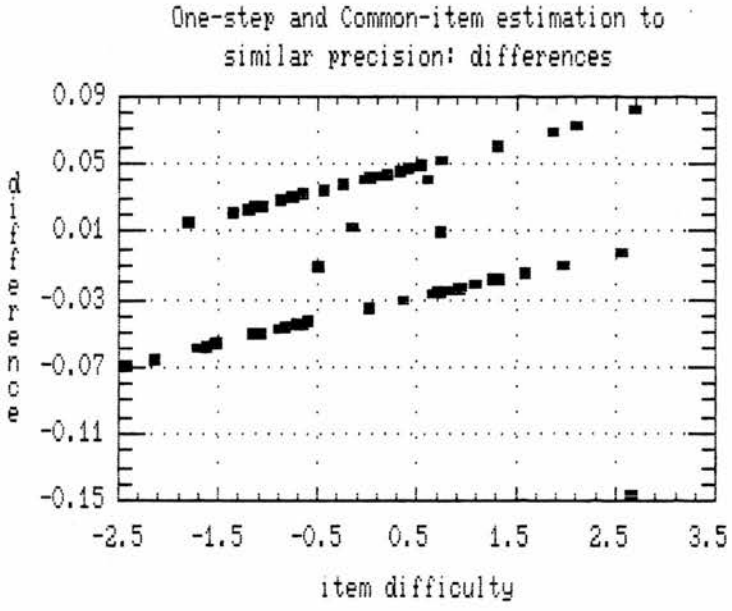


Fig. 4. Difference between one-step, common-item estimations (tests 28, 38)

Fig. 4 can be interpreted to show that the one-step approach is superior in two respects:

1) It separates items better, i.e. produces a longer scale. The harder items are harder, the easy items easier, in the one-step estimation: hence the general slope of the plot.

2) It separates the tests better. The upper line is the harder test, the lower line the easier. The fact that they are clearly separated shows that the one-step estimation has improved on the values found by common-item methods.

Curiously, these two effects work against each other, with the result that the easier items in the harder test, and the harder items in the easier test, finish up around the zero mark; that is, they are similarly estimated by both approaches. It is the hard items in the harder test, the easy items in the easier test, which benefit most from the one-step estimation. But again, the actual differences are rather small.

This small study suggests that with the algorithm used, further iteration would have little effect, and thus that the difficulties found can be considered stable. However, this finding does not agree with Lee (1991). If the reason for this is in the algorithm, one possible factor is the stopping value for the inner Newton-Raphson loop (which is very loose in the present algorithm). This study might be repeated with different values.

6: Explaining item difficulty

The previous chapter focussed on the practical problems of constructing a trait by calibrating items and fitting them to the common bank scale. The latter part of that chapter described a number of attempts to relate quality of fit to such factors as particular criteria for correctness, item type, item difficulty, item content and the influence of L1 transfer. Any insights gained from such investigations contribute to a picture of what it is that the bank trait is really measuring; that is, they are a part of construct validation. This chapter takes construct validation further by asking the question: 'What makes items difficult?' Both quantitive (multiple regression) and qualitative analyses are undertaken in an attempt to answer this question.

6.2 Causal models, levels of abstraction

The set of test items, calibrated for difficulty, are the raw material for the second stage of the construct validation study: an attempt to explain why it is that some items are harder than others, and thus to say what it is that the test measures. The notion of 'explanation' implies doing more with this raw material than simply subjecting it to exhaustive quantitative analysis. Explanation implies the construction of a theory. As Blalock puts it:

> One can readily point to the possibility of assembling so many miscellaneous facts on a subject that it becomes virtually impossible to make any sense out of them. But empirically-minded quantitative sociologists sometimes in effect endorse an anti-theoretical position by throwing numerous variables into a regression equation with the idea of selecting out that subset which 'explains' the most variance. (Blalock 1969:2)

As Blalock points out, such an approach may be satisfactory in particular situations where simple prediction is the goal; but it is unsatisfactory if we would wish to be able to generalize to new situations where certain factors are different. Then it becomes necessary to understand the causal relationships between factors, as well as the possibly complex forms of interaction between them. This is where theory becomes necessary.

The notion of causality is thus fundamental to theory building, and any attempt at explanation must state causal relationships, at the risk of course of being proved wrong.

> The basic strategy of the analysis of causal models is first to state a theory in terms of the variables that are involved and, quite explicitly, of what causes what and what does not.... The observational data are then employed to determine whether the causal model is consistent with them, and estimate the strength of the causal parameters. Failure of the model to fit the data results in its falsification, while a good fit allows the model to survive, but not be proven, since other models might provide equal or better fits (Cohen & Cohen 1983:14).

This point is worth stressing, given the strength of the popular conviction that 'correlation does not mean causation', and the consequent tendency to treat abstinence from imputing causation as a positive virtue. Cohen & Cohen find the above saying 'although well intentioned, to be grossly misleading. Causation manifests itself in correlation, and its analysis can only proceed through the systematic analysis of correlation and regression' (1983:15).

Although we would wish to state a theory of test difficulty in terms which might have some generalizable meaning, we have to agree with Blalock that 'it may be very difficult to formulate highly general theories that imply predictions taking us very far beyond the common-sense level of analysis' (Blalock 1969:141). That is, there is evidently a tension between the desire to generalize on the one hand, and the possibility of demonstrating detailed, interesting relationships on the other.

Any theory *can* only be tested by applying it to data, which means that abstract concepts in the theory must at some stage be linked with *indicators* that can actually be measured. This may involve making 'certain a priori untestable assumptions concerning the causal linkages involved' (Blalock 1969:151). This distinction between *variables* and *indicators*, that is, between concepts which have some generalized theoretical status, and concrete data which are believed to constitute measures of them in a particular instance, seems to be an important one which is not always explicitly drawn in language testing research.

Imputing causal relationships to variables in a theory implies putting them in some logical order, as well as specifying the nature of the relationships between them. Language testing research to date appears to restrict itself to positing simple additive relationships among variables: that is, where several variables are held to bear on difficulty, their total effect is expected to be more or less the sum of their unique individual effects. This may be a reasonable assumption, given that 'in general, additive models seem to approximate reality reasonably well' (Blalock 1969:156); but the possibility of non-additive (interaction) effects might also be entertained.

6.3 A model for test item difficulty

Having sketched the parts of a causal model in general terms, let us attempt to outline a model for the test items in the present study.

It is important to start at the beginning, that is, with the variables that we consider logically prior to all others. But it is by no means clear where the beginning is. Pollitt & Hutchinson (1986) begin their account of the question answering process at the moment the examinee begins to read the paper; and yet this might equally well be seen as the *end* of a process which began possibly years before with a test constructor sitting down to produce a test specification, and an item writer producing items to this specification. Actually it is not quite the end, if you take marking the papers to be a logically as well as temporally subsequent stage in the testing process; although a marking scheme and criteria for correctness will also (in all probability) have been established before anybody sits down to take the test.

This may seem a trivial point, and yet it is of practical importance for the results of a multiple regression analysis, as the order in which variables are added (which corresponds to the causal order proposed by our model) may well have a bearing on *which* variables appear to account for item difficulty. In the language of multiple regression/correlation (MRC) analysis: where two or more independent variables are correlated with each other, as well as with the dependent (Y) variable, the one which is entered into the equation first will be seen to account for a proportion of Y variance (test item difficulty); but those entered subsequently, to the extent that they co-vary with the prior variable, will not account for any more Y variance.

6.3.2 The language problem is logically prior

Now the logical beginning for explaining an item's difficulty is surely to ask: 'What is the item about?' - that is, what was the item writer's intention in writing that item? Considerations of *how* the examinee is induced to demonstrate knowledge of this problem are logically subsequent to this.

This distinction between the 'what' and the 'how' of language testing is difficult to draw unambiguously, given that 'language is both the instrument and the object of measurement' (Bachman 1991:2), and yet it is surely an important distinction to maintain, as the following treatment will attempt to show.

Bachman provides a taxonomy of *test method facets*, which

> constitute the 'how' of language testing, and are of particular importance for designing, developing and using language tests, since it is these over which we potentially have some control (Bachman 1990:111).

Further on he underlines that method facets are essentially extraneous to the language ability being tested:

> The effects of ... the methods used in language tests may reduce the effect on test performance of the language abilities we want to measure. ... For this reason, it is important to understand not only the nature and extent of these effects, but also to control or minimize them. (p.156)

But Bachman's taxonomy in fact reads more like a comprehensive framework for describing language tests both with regard to content and method. Indeed, he advocates its use for the comparative description of language tests (p.152), and his report of research done as part of the Cambridge-TOEFL Comparability Study seems to be as much about content as method.

What is problematic about the taxonomy is perhaps that it puts on an equal footing factors of unequal status. There are aspects of tests which are clearly peripheral, 'method' factors (e.g. the way a test is divided into parts or sequenced, the way the instructions are given); aspects which *define* kinds of language test (e.g. of different skills - listening, speaking etc.); and aspects (under the heading *Nature of language*) which are essentially linguistic and thus inextricably bound up with content.

Whether some aspect of language is to be seen as a method facet or as central to what the test is about will depend on the purpose of the test (as Bachman himself states). The difficulty of vocabulary used, for example, might be considered a method facet in an oral interview, whereas it would certainly constitute the content of a vocabulary test. The Item Banker items are in the main about grammar, which appears in the taxonomy under the sub-heading *Organizational characteristics*. It will not be useful to treat grammar as a method facet whose effect is to be 'controlled or minimized'.

So although a taxonomy such as Bachman's may serve the purposes of *description*, any attempt to *explain* the difficulty of items must start from considering what the item is really about, which is to say that it must acknowledge the principle of causal order.

An example will illustrate further the difference between description and explanation. Let us take the distinction between selected and constructed response item types (Popham 1978). Bachman (1990:129) cites research that 'supports the intuitive hypothesis that constructed response types will generally be more difficult than selected response types.' Yet lacking further qualification this statement is evidently false. There are a range of examinations that manage to achieve a high level of difficulty entirely within a multiple-choice (selected response) format - the TOEFL, for example. At the same time we know that constructed response types are appropriate for exercises and

classroom tests at beginner level. Clearly, the difficulty of an item reflects above all the intention of the item writer, who having selected a language problem with the desired degree of difficulty can devise a form of presentation using whatever response type is required.

Thus to establish that a test consists, say, entirely of selected response items, though useful for descriptive purposes, can explain nothing about the difficulty of those items. If, as in the case of the Item Banker items, both selected and constructed response types are represented, it is of no interest to investigate whether one type is easier than the other - or rather: it may be of descriptive interest, but it cannot serve to explain anything.

This is no less true of various other structural features which one might, with more or less intuitive justification, wish to associate with item difficulty - the number of words or clauses in the stem, for example, the number of words to be written in the response, the position of a gap, or the part of speech of a gapped word, etc. Attempts to associate item difficulty with such structural features (e.g. Curtis 1987, Freedle & Fellbaum 1987, Fellbaum 1987) tend to show modest or zero correlations, and anyway, even if substantial correlations were ever to be found, their interpretation would remain problematical, in the absence of some explicit treatment of what each item is about.

To return to Bachman's contention that 'constructed response types will generally be more difficult than selected response types.' The implied condition is: 'other things being equal'. That is, given two items about *the same language problem*, the constructed response will be more difficult than the selected response type. In the language of causal order again: the language problem variable is prior to the constructed/selected response variable. We have to *control for* the language problem before examining the effect of method factors.

Thus we first need to operationalize the notion of a *language problem*, and investigate its influence on item difficulty, before attempting to

assess the influence of structural features of items such as response type, number of words to be written, difficulty of vocabulary employed, and so on.

The notion of the Language Problem will be refined below (6.4). For now we take it to be the starting point of our model of item difficulty, and attempt to move on from there.

6.3.3 Rubric

We now arrive at the moment when the examinee picks up the test paper and begins to read the first question. The first thing to read is the *rubric*. As a variable in the analysis we use the term in a wider sense than in Bachman's taxonomy, to denote not only the specific instructions given to the examinee (a written text at the head of each group of items, in the present case), but also what Bachman calls *facets of the expected response* (for example, the type of response - selected or constructed). We treat these together because they are, unfortunately, impossible to separate. Although we would wish to investigate separately the effect of the wording of the instructions (it might be ambiguous, misleading or confusing, for example), we cannot do so, at least in the quantitive part of the present study, because the rubric-as-text and the rubric-as-task identify exactly the same group of items. Whatever influence the text of the instructions may have on item difficulty thus cannot be separated quantitively from the influence of the task that they specify. This also presents a problem for the causal model: where should the rubric variable be placed? It can only go into the model once, and the best solution seems to be to put it later on, because the effects of the rubric-as-task are of greater interest than the rubric-as-text.

The qualitative analysis of groups of items (below,6.4.4) promises to offer more insight into the influence of the rubric-as-text.

6.3.4 The Prompt

Having read the rubric the examinee proceeds to read the body of the item: that is, all the text given on the page, including the given parts of gapped or incomplete sentences. Grammatical or lexical difficulty will make itself felt here. Indicators of grammatical difficulty might include: lexical density, i.e. the ratio of lexical words to total words; and the noun/verb ratio. There are many possible indicators of lexical difficulty. Perkins & Linnville (1987) found significant predictors of difficulty of items in a vocabulary test to include: word frequency, number of syllables, number of letters, abstractness, distribution. Some of these indicators seem to exemplify the 'a priori untestable assumptions' about causal linkage mentioned above. Number of syllables and number of letters, for example, are held to be indirect measures of a word's pronounceability, it being argued that 'a word which is easily pronounced is less difficult' (Perkins & Linnville 1987:133).

Reading the text the examinee creates what Kay (1987) calls an 'envisionment': a mental representation of the events or situations described in the text, and of the characters and other participant roles explicitly mentioned or implied. In doing so he uses 'conventional or stereotypic representations of "knowledge of the world" as a basis for the interpretation of discourse' (Brown & Yule 1983:236). Brown & Yule review work in AI (artificial intelligence) and psychology on such representations: they discuss *frames*, *scripts*, *scenarios*, *schemata* and *mental models*, terms they consider as 'alternative metaphors for the description of how knowledge of the world is organised in human memory, and also how it is activated in the process of discourse understanding' (p. 238). A criticism frequently levelled at short discrete items of the type used in Item Banker is that they elicit unnatural, unrepresentative language performance because they lack any context. Of relevance here is the long-running debate as to the amount of context which is made use of in doing cloze tests, Porter's (1983) conclusion being that context beyond 5 or 6 words is *not* used. While the lack of textual context is an undoubted limitation of discrete items, they certainly do have

context in the sense of an evoked mental representation, because the reader *supplies* the context in the process of making sense of the text. Brown & Yule (1983) stress that discourse is not constituted by so much by particular features of a text, but rather by what speakers and listeners bring to the interpretation of text.

This has implications for what features of items might be connected with difficulty, as well as what features might *not* affect difficulty. Items which evoke more familiar, stereotypical schemata should be easier. Texts with more concrete reference (narrative, description, dialogue) should be easier than more abstract texts (e.g. containing logical arguments. On the other hand, we might expect the presence or absence of explicit linguistic markers of textual or logical relations to make little difference to difficulty. It is not in fact easy to apply these indicators unambiguously to items in the present study (largely because of the minimal textual context). But the group of items containing jumbled sentences provide enough text to categorize as being narrative or not; and other items can with more or less confidence be categorized as representing dialogues.

Having created the 'envisionment' the examinee then has to infer the meaning which has to be conveyed in the response (note that language tests of this type work by prescribing meanings in order to constrain particular linguistic responses).

6.3.5 The Response

Next comes the composition of a response. There is a basic split here between selected-response and constructed-response items, and some smaller splits between different types of selected-response item: that is, it is not possible to apply all indicators to all items.

Constructed-response types

There is most to say about constructed-response types.

Firstly, strength of the *elicitation*. The explicitness with which a particular response is triggered is something over which the item writer has considerable control, and consitutes a mechanism whereby more or less *support* (Pollitt & Hutchinson 1986) can be offered to the learner. On the other hand, elicitation can go wrong in a number of ways. *Miscueing* is evident when a disproportionate number of learners provide the same, incorrect response. An example:

They prefer to stay at home; they go hardly _____.

This is a very difficult item, partly because the parallelism of 'staying at home' and 'going out' leads most learners to respond: 'they go hardly out.' Problems with elicitation often spring from the item writer's failure to elicit the intended response strongly enough, leaving the learner hunting around for some other possible response. Where such an unforeseen response exists and is easier (a *getout*, let us say) the item becomes easier than it was meant to be; where no getout exists, the item becomes harder. Elicitation may sometimes depend on conventional or cultural knowledge which may be thought ancillary to what the item is meant to be about. Two examples:

Porter: Here are your bags.
Lady: You ..... a mistake. These are not our bags.

Newsreader: Here is the news.
There / be / earthquake / Japan

Both these items are intended to elicit the Present Perfect, but the reason that 'You are making a mistake' is unacceptable as a way of pointing out an error lies not so much in grammar as in custom. The unacceptability of 'There is an earthquake in Japan' seems to lie in the fact that earthquakes are normally perceived as completed events rather than continuing states, in contrast, say, to government crises, forest fires or wars. This again seems to be more a matter of convention than grammar. Whether or

not this makes these items any worse as items, it certainly makes them unrepresentatively difficult for the language problem they were devised to test.

Indicators for the elicitation variable might include: how the response is immediately constrained by text (on the right, left or both); whether the prompt might miscue; whether it depends on cultural or world knowledge; whether a unique response was intended; whether other responses are possible; whether a getout exists.

The *size and nature of the task* might be taken as the next variable (although they follow fairly directly from the elicitation). The number of words to write, and the number of lexical words to supply, the number of words copyable from the given text, are fairly direct indicators of one kind of difficulty. A task can be called *mechanical* if it calls for no more than the application of a rule in response to an explicit instruction. It can be called *paradigmatic* if it calls only on knowledge of closed word classes or grammatical rules (excluding idioms, or lexically-determined rules such as verb complementation). Paradigmatic and mechanical tasks are expected to be easier. The lexical difficulty of the response can be measured by the same indicators as the prompt.

'Writing down the response' might be included as the next stage, although it does not seem particularly productive, as the main factor, spelling, seems better treated under the following heading.

Selected response types

The various selected-response types used in Item Banker offer different kinds of difficulty at the response stage. In particular, jumbled words items can be expected to be more

difficult the more words to unjumble there are. Where there is a genuine word-ordering problem involved (such as the position of adverbs) we can expect this to add extra difficulty.

The feature of being a narrative text is one that can properly only apply to the jumbled-sentences item type; so this indicator is reserved to this particular selected-response type.

6.3.6 Marking the paper

Marking the paper is temporally and logically the final stage in the model of item difficulty. Evidently the criteria for giving or witholding marks can make an enormous difference to the effective difficulty of an item. Decisions on what kinds of spelling mistakes to allow, whether to tolerate minor mistakes in non-essential parts of the response, or which particular unforeseen responses to accept, will alter item difficulty drastically, and not uniformly across all items (as different items attract different kinds of mistake). As suggested above in the discussion of model fit, there is much of potential interest here, although a factor militating against following this up in the present study is the time-consuming nature of re-scoring papers.

6.4 A Multiple-Regression analysis

The sequential model introduced above is summarised in Table 6.1. It can be given explicit expression in the form of a multiple regression/correlation (MRC) analysis, as shown in Figure 6.1. This bar-chart shows the item features believed to be associated with item difficulty (the indicators) listed on the Y-axis in the order they are introduced into the analysis. The horizontal bars represent the total amount of variance in item difficulty accounted for after introducing each indicator. Thus the importance of each indicator is shown by the

	VARIABLES	INDICATORS
Before the test:		
1	Choice of the language problem	
During the test:		
	(Reading the rubric)	
2	Reading the prompt	
3	Grammatical difficulty	a) lexical density b) noun/verb ratio
4	Lexical difficulty	c) frequency of hardest word d) mean frequency e) mean lexical frequency f) freq of selected key word
5	Creating an envisionment	g) text is narrative? h) text is dialogue?
6	Inferring meaning	i) negative? j) counter-factual? k) modal?
7	Responding	
8	The elicitation	l) constraint: left/right/both m) is misleading? n) cultural knowledge etc? o) unique response expected? p) other responses possible? q) getout possible?
9	The size of the task	r) no. of words to write s) no. of words copyable t) no. of lexical words u) task is mechanical? v) task is paradigmatic?
10	Lexical difficulty	(as c - f above)
After the test:		
11	Marking the paper	

Table 6.1 A sequential model of the item-answering process: Variables and some possible indicators

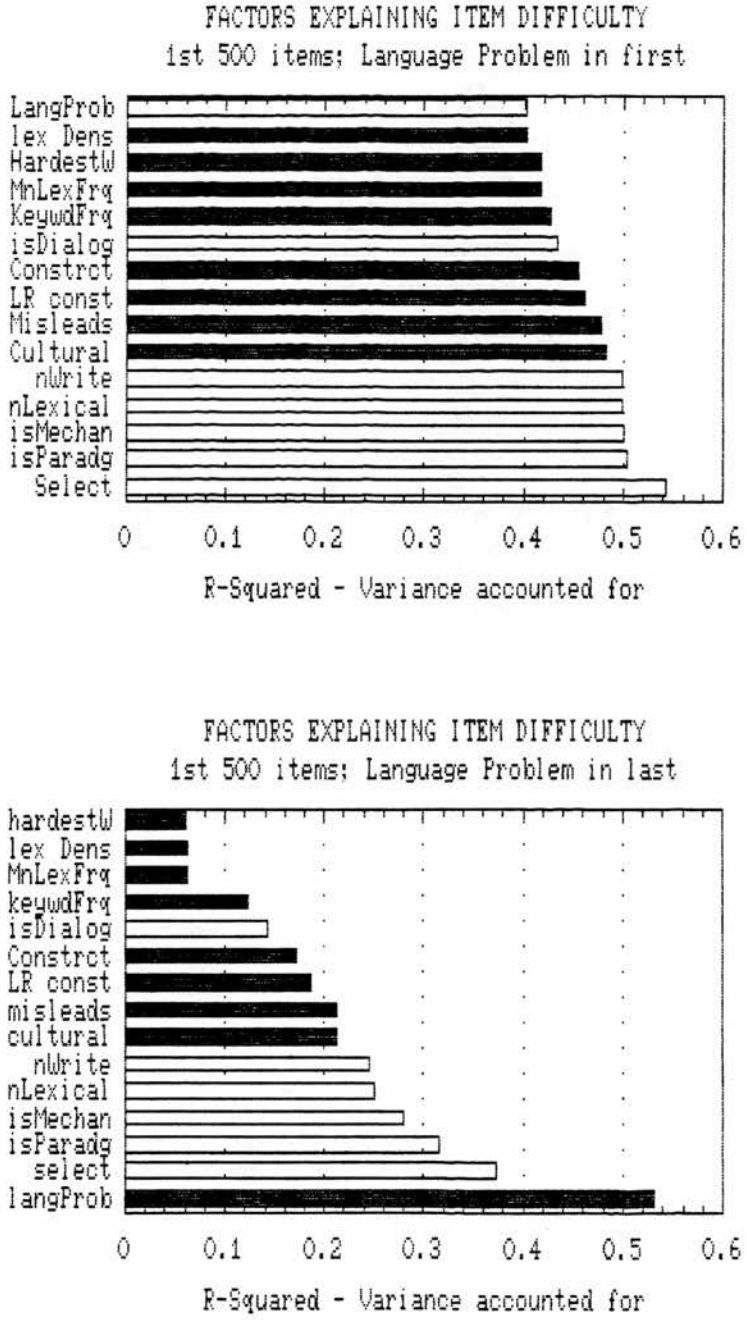


Figure 6.1 Two causal models of item difficulty (1st 500 items). Language Problem is a) entered first b) entered last.

difference in length between one bar and the next. The X-axis shows the proportion of variance accounted for: the analysis in Figure 6.1. accounts for just over half of item difficulty.

It is important to realize that a MRC analysis does not prove, disprove, or do anything to indicate the superiority of one causal theory over another. The total power of any set of indicators to predict the dependent variable (here, item difficulty) will be the same, whatever order they are put into an analysis. To the extent that different models introduce different indicators, one might claim that the one which turns out to have more predictive power is to be preferred (but recall Blalock's warning that predictive power is not the same thing as explanation). What a MRC analysis does is to make explicit the consequences of adopting a particular causal theory. The superiority of the theory is demonstrated not by the analysis but by the logical arguments put forward in its support.

Figure 6.1, showing two different MRC analyses, illustrates two competing theories of item difficulty. The first (a) is the theory advanced above, i.e. that the language problem is logically prior, and must be introduced into the analysis first. The second (b) shows the consequences of leaving the language problem out of account until all other factors have been given a chance. Thus the two models differ only in the placing of the language problem variable: it is first in (a), last in (b). The order of the other indicators, which corresponds to the hypothesized question-answering sequence described above, remains the same in both models (an explanation of the labels used will follow directly).

If one accepts the argument advanced here, that the language problem is logically prior, that is, that the difficulty of a language test item resides chiefly in the selection of the language problem by the item writer, the consequences for explaining item difficulty are clearly shown in Figure 6.1. The language problem accounts for 40% of item difficulty (in an analysis of the first 500 items). Many of the other factors add no explanatory power, and all of them together

account for little more than 10% of item difficulty. In this view, it is the 'what' of language testing rather than the 'how' which is decisive: the significance of the 'test method facets' discussed by Bachman (1990) appears greatly diminished.

Contrast this with the alternative model in which the language problem is left out of account. Method facets alone account for over a third of item difficulty. If one accepts this view, they remain an interesting subject for further research.

Let us now look in more detail at the other indicators included in the two models.

First come four putative indicators of grammatical and lexical difficulty, associated with the *Reading the prompt* variable: lexical density, the hardest (least frequent) word in the text, the mean frequency of lexical words in the text, and the frequency of one word selected as being central to understanding the text. In the analysed data set lexical density accounts for nothing. The remaining indicators are just some of a variety of attempts to associate vocabulary difficulty, operationalized as word frequency, with item difficulty. It is striking that the association remains rather weak, however many ways one attempts to measure it.

Next comes the indicator *is a dialogue?*, which is supposed to relate to that stage of reading the prompt where the testee creates an 'envisionment'.

Next come four indicators of difficulty associated with the elicitation of the response: whether the item is constructed response, how many immediate textual constraints there are, whether the prompt might mislead, and whether it invokes cultural knowledge.

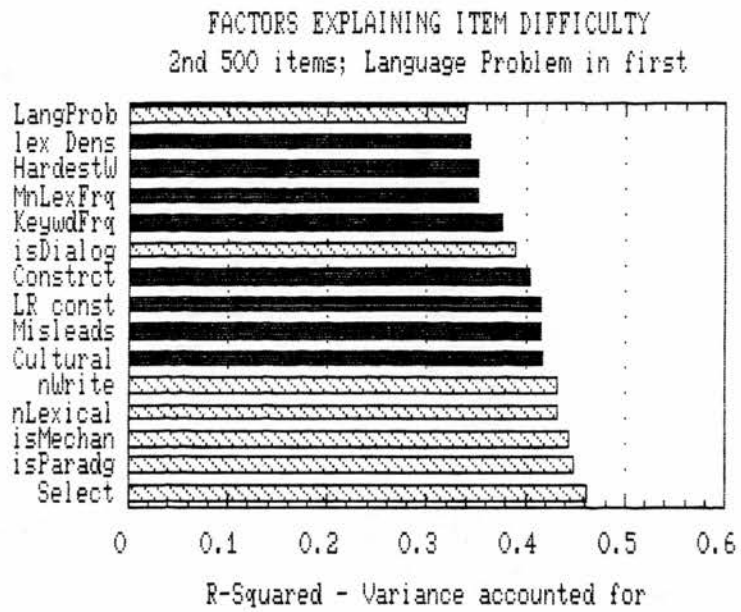
Next come four indicators of difficulty associated with the size of the task (in constructed-response items): the number of words to write, the number of lexical items to supply, and whether the task is mechanical or paradigmatic.

Lastly comes one variable which summarizes a number of indicators applicable to selected-response items, as outlined above (6.2.4).

It may be observed that the analyses illustrated here do not incorporate all the indicators in the model of the item-answering process outlined above. This is because the analyses chosen for presentation here include only a 'shortlist' of the more promising indicators, many others having been discarded earlier. Strictly, this is no way to test a causal model. The model should be constructed first, and only then tested out in a MRC analysis. The dangers of the trial-and-error approach adopted here, sometimes called 'heuristic', and elsewhere 'data snooping', should be clearly admitted. By accepting those variables that 'work' into the model, and rejecting those that do not, chance is capitalised on, and the significance of signficance tests is therefore undermined. As a result, there is no statistical basis for generalizing from findings in the present study to other data. We might expect the model to have less predictive power when applied to fresh data, and Figure 6.2 shows that this is indeed the case.

Figure 6.2 shows the same two models applied to an analysis of the second 500 items. The same general picture emerges, but overall, the models account for a smaller proportion of item difficulty (about 46%, as opposed to 54%).

Fortunately the argument presented here does not depend on demonstrating the significance of all the variables in the model, but rather the secondary importance of all variables relative to that of the Language Problem variable. This is clearly enough shown, both in 6.1 and 6.2. What does need explaining is how the language problem variable was operationalised in these analyses. To use the language problem as a variable, items were grouped by language problem and each item given a value which is the mean difficulty of the group. This



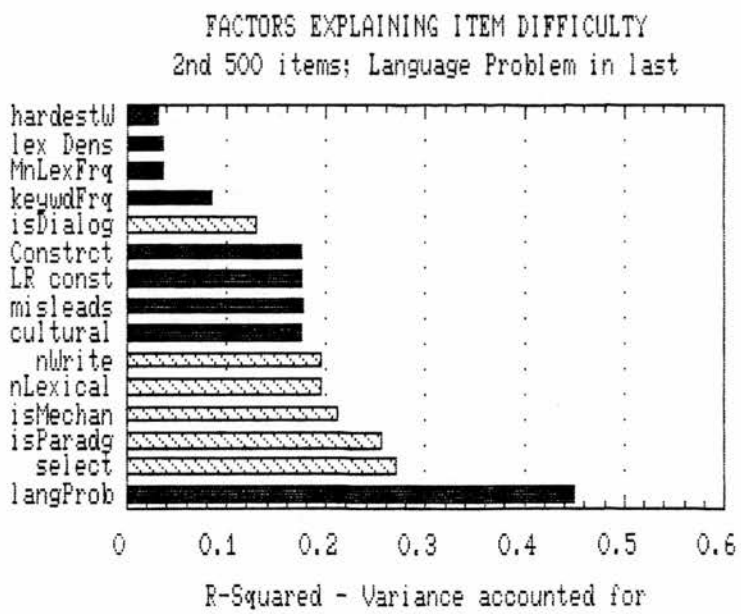


Figure 6.2 The two models of Fig. 6.1 applied to 2nd 500 items. Language Problem is a) entered first b) entered last

of course looks like feeding back part of Y variance into an X predictor, and it may not seem surprising that it accounts for a large part of Y; but a similar result would be obtained simply by dummy-coding language-problem group membership.

6.5 A qualitative analysis of groups of items

The sort of quantitative analysis described above turns out to be of rather limited value in understanding what features account for the difficulty of particular items. This is simply because there are not many features which it makes sense to apply to all items. More interesting and interpretable results are obtained when one looks at smaller sets of items and investigates the effect of features which are relevant to those items. Generally, a problem for the interpretation of correlational analyses like MRC is that a 'significant' effect - that is, an improvement in R-squared due to a particular indicator - is a portmanteau value that represents both the strength of the effect on the items to which it applies, as well as the proportion of items to which it does apply. If a feature like vocabulary difficulty is found to bear a weak positive relationship to the difficulty of items, nothing about the analysis will reveal whether this is because vocabulary difficulty is very important in the case of a few items, or slightly important in the case of all items.

Many features cannot be sensibly applied to all items. The number of words to unjumble, for example, is a feature which can only apply to the jumbled-words item type. The MRC analyses described here managed to keep all the items together by coding inapplicable cases as missing data, as explained by Cohen & Cohen (1983). This has no effect on the total amount of difficulty accounted for, but neither does it make for clarity: the apparent significance of each feature is watered down.

The case of lexical difficulty may be taken as an example of the problem with mechanical, comprehensive analyses. As mentioned above, various attempts were made to trace an association between lexical difficulty and the difficulty of the item. This seemed like an

attractive idea initially, as lexical difficulty could be operationalised fairly readily, using a word frequency list supplied by Cobuild. The list was already on computer disk, and software was written to confront each item's text with the frequency list, and generate various measures quite automatically.

The results were disappointing. The measures correlated with difficulty not at all, or only weakly. The strongest association was the least frequent (hardest) word in the item ($r = -.20$). Yet it seems unlikely that lexical difficulty should be so weakly associated with the difficulty of language test items. Part of the problem is certainly the over-simplifying operationalisation of lexical difficulty as word frequency, but there is more to it than that. The fact is that difficult items can be framed in simple words, and vice versa. It is necessary to start looking at the items.

So the next stage was to select those items where a 'key word' appears to be particularly relevant to responding correctly. Less than half the items qualify for this group. The frequency of the key word correlates substantially higher with difficulty ($r = -.46$).

Finally, taking the group of items with the rubric 'Complete the sentence with the correct form of the word in brackets' there was a high correlation between the frequency of the word to be supplied and the item's difficulty ($r = -.71$).

This example shows that quantitative measures can be more revealing when applied to relevant data; and identifying what is relevant requires qualitative analysis. The MRC analysis described above has been interpreted to show that when the *language problem* is correctly located as the first causal factor in item difficulty, then there is little to be gained by looking at the remaining structural features of items. Perhaps it would be truer to say that there is little to be gained by attempting brute quantitative analysis of the whole item set using one blanket list of features. In fact it is extremely revealing to examine items, *after* grouping them by language problem. Effects become evident which would otherwise escape notice.

We have been invoking the language problem for some while; now it is time to provide a definition of this notion.

6.5.2 The 'Language Problem' defined

In fact, the definition of a language problem need not be theoretically rigorous. Any categorization of language will serve as long as it allows us to group items into sets which are roughly about the same thing - sufficiently so, that is, for it to be possible to study items together and compare them with each other. The following discussion is based on analyses of groups of items selected from the Item Banker database by their content tags, and these tags mostly represent traditional pedagogic categories: 'Present Perfect', 'Passive', 'First Conditional' etc. A language problem obviously can overlap with or subsume other language problems.

We have argued that it is necessary to model the *inherent difficulty* of language problems, and provided some quantitative evidence that this is the major factor responsible for the difficulty of language test items. Only having done this can we hope to disentangle the contribution to difficulty made by other, structural features of items.

As a first approach to a theory we might propose that each language problem has a *true difficulty* (by analogy with the notion of a testee's *true score* in classical test theory). Just as the true score is that score a person would obtain if all measurement error could be removed, the true difficulty of a language problem is that difficulty rating it would receive if all method effects could be neutralised. We can conceive of method effects that would make an item unduly easy, providing too little challenge or too much support - 'giving the answer away'. Many effects certainly make items unduly difficult: ambiguity of instructions, misleading cues, extraneous vocabulary difficulty, under-elicitation (that is, a prompt that is too indeterminate to evoke the intended response), etc.

This proposal is unsatisfactory because second-language acquisition research has already demolished the idea of language form being composed of unitary entities that are somehow acquired whole (which is what having a single difficulty rating would entail). Recall the criticisms of the Morpheme Studies rehearsed in an earlier chapter (2.2.3). Mastery of some aspect of language form proceeds along a number of gradients; of relevance to this discussion are: the progression from receptive to productive use; from holistic, formulaic use in very familiar contexts to analytic use in unfamiliar contexts; from use with a single functional meaning to use with a variety of functional meanings, and so on. If we are to capture these developmental dimensions in our characterization of item difficulty, then we should not try and marginalize them as 'method facets'.

So perhaps a more appropriate metaphor is of a *difficulty envelope* - a range of difficulties associated with different stages in the acquisition of a language feature. The lower end of the envelope would relate typically to familiar, formulaic, concrete contexts, and the higher end to more creative, abstract or perhaps literary contexts.

I believe it is possible to draw a workable distinction between features of test items which reflect the different contexts of use and meaning constituting the 'true' difficulty envelope, and features which are genuinely extraneous, method factors. How one attempts to draw such a distinction will tend to depend on principles of descriptive economy. A feature which applies equally to all language problems will be best considered a method facet. Thus if it is true of all language problems that receptive ability precedes productive use, and thus that items testing reception will be regularly easier than items testing production, then this distinction (essentially the selected/constructed distinction again) can be considered a method facet. Doing so will allow us to narrow the range of the difficulty envelope.

It is relatively easy to identify extraneous method facets which make items unduly difficult. Some have been mentioned already. One additional problem which should be pointed out is the case when an item intended to be about one particular language problem contains another more difficult language problem. From the point of view of the first problem, the second one is an extraneous complication. When studying a group of items on a given language problem it may be difficult to judge where to fix the upper end of the difficulty envelope, because the harder items tend to merge into other language problems.

It is also difficult to find a principled way of fixing the lower end of a language problem's difficulty envelope. Many items offer so much help that a correct response in no way demonstrates mastery of the tested problem. A correct response to the gapfill item:

_____ is your name?

(which happens to be the easiest item in the bank) certainly does not demonstrate mastery of WH- question formation. Thus the WH- language problem is not as easy as this item alone would suggest. A learner who gets this item right and other more difficult items wrong is demonstrating some partial knowledge of the problem. It would be too much to suggest that such items can throw much light on developmental processes in second-language acquisition terms - the fact that all items demand accuracy, and no points are awarded for interlanguage forms, makes it clear that although such items test partial knowledge, they can not be taken to reflect directly the state of some underlying transitional competence. None the less, such items may well be reliable indicators of a learner's overall ability level, and it may well be useful (e.g. for teachers or materials writers) to know what kind of tasks learners will typically be able to perform having only partial knowledge of a language problem.

Thus we may choose to include such items within the profile of a given language problem, while indicating that they demonstrate something less than mastery. This would mean fixing a threshold

point at the transition from pre-mastery to minimal mastery. A workable criterion for fixing this point would be that items at or above it should be constructed-response type, requiring the learner to supply a reasonably complete instantiation of the given language problem.

6.5.3 Examples from the bank

Figures 6.3, 6.4 and 6.5 illustrate profiles of item difficulty for 12 areas of grammar, based on analysis of groups of items selected by their content tags. Each numbered heading groups one or more sets of items, identified by a label on the left. Thus heading 1), Question Formation, has difficulty envelopes for DO-questions, question formation by inversion, and question tags. These are shown against a nine-point scale.

A broken line to the left of the envelope indicates pre-mastery - that is, items that are too easy, offering too much help to indicate true mastery. Sometimes there is a broken line to the right, indicating areas where it is difficult to assign a clear upper limit to the difficulty envelope. Thus under heading 3) the envelope for the Past Simple is extended right along the scale, encompassing the aspect of lexical difficulty. Here, and in other places where it seems illuminating to do so, lexical items or other glosses have been added at appropriate points on the scale.

These figures give an impression of what might be achieved in time through an item banking approach: a detailed, explicit picture of how language proficiency develops. The present study is no more than a start in this direction, and this should be borne in mind when looking at these figures.

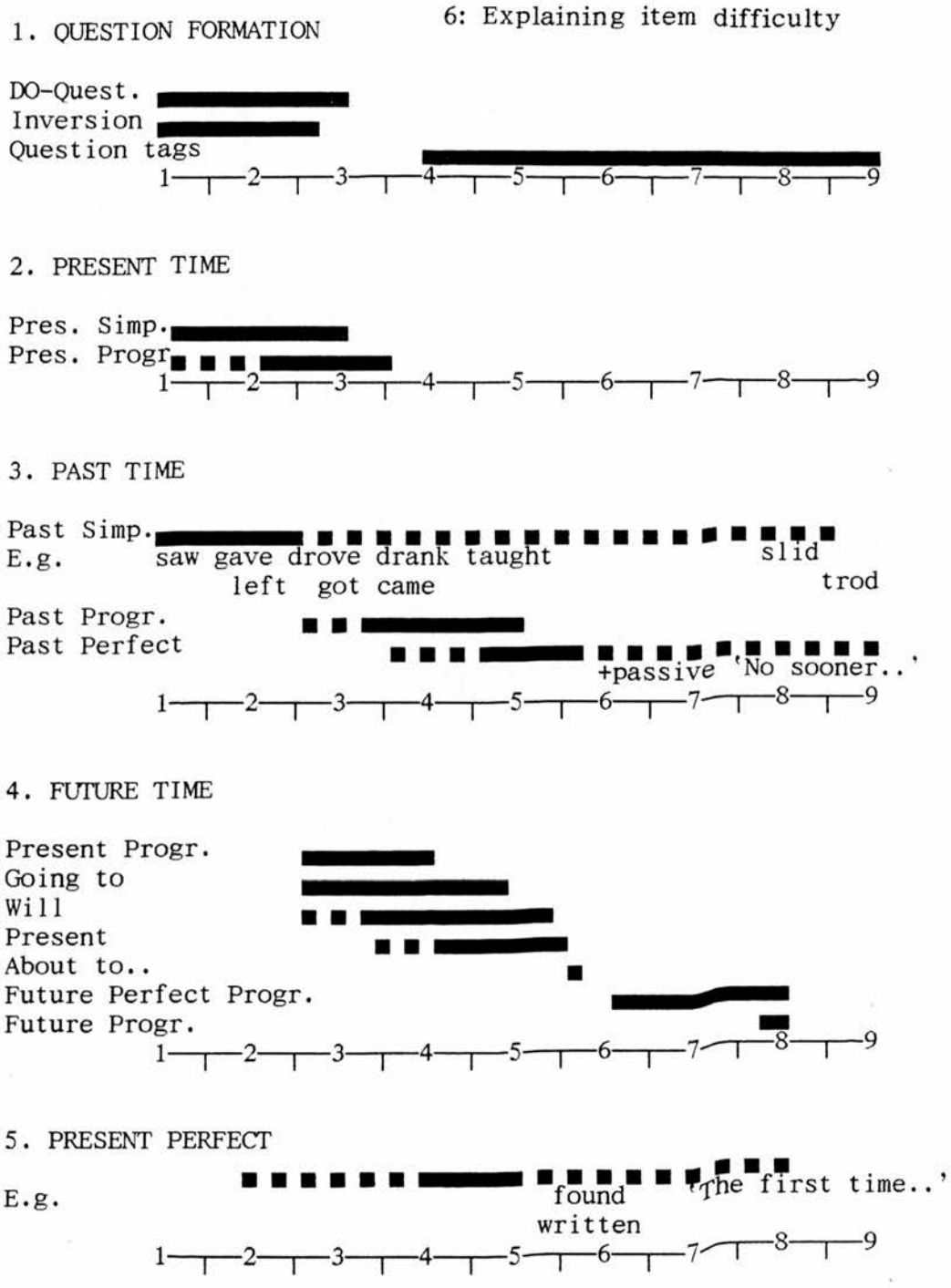


Figure 6.3 Examples of Difficulty envelopes (1)

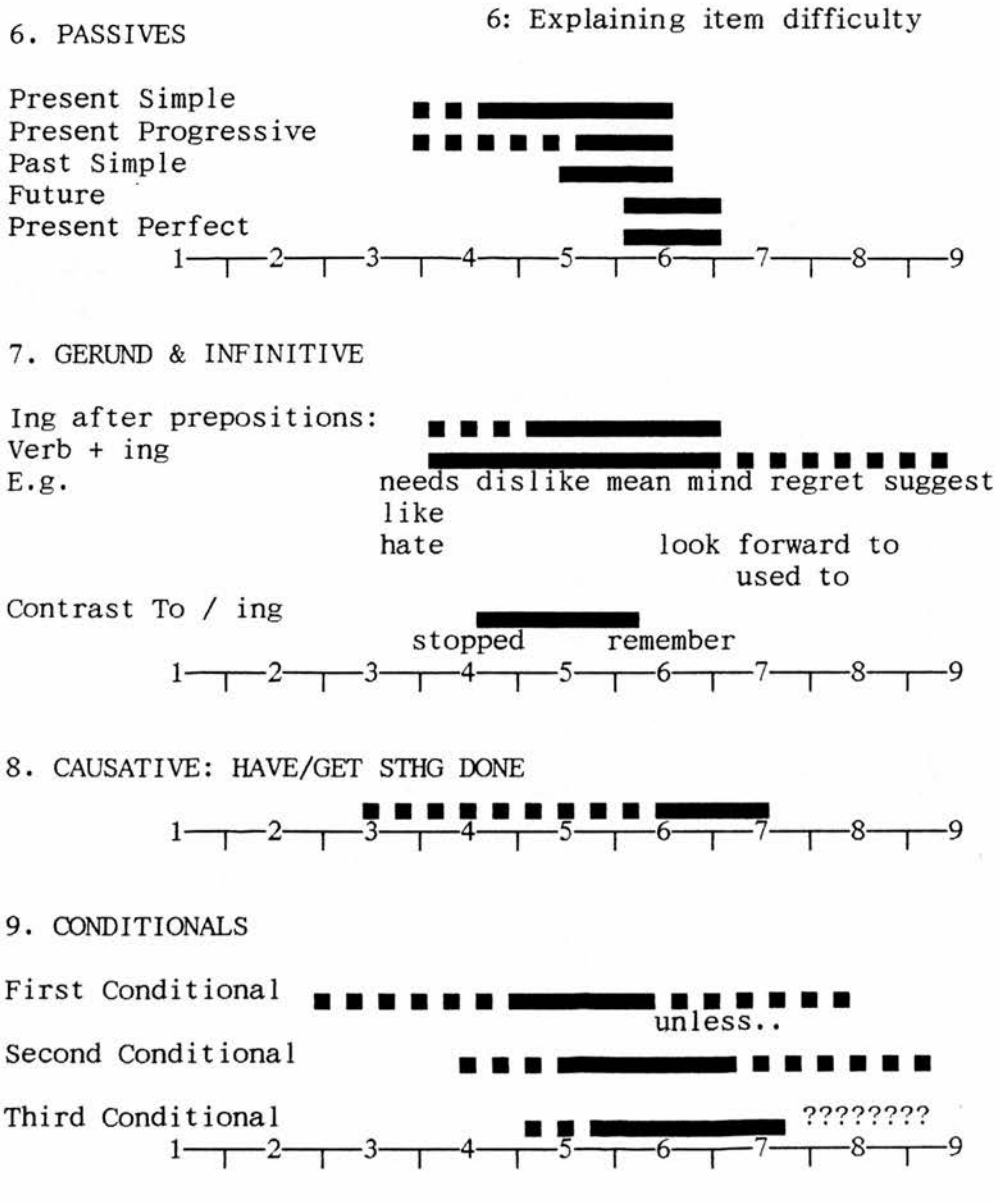


Figure 6.4 Examples of Difficulty envelopes (2)

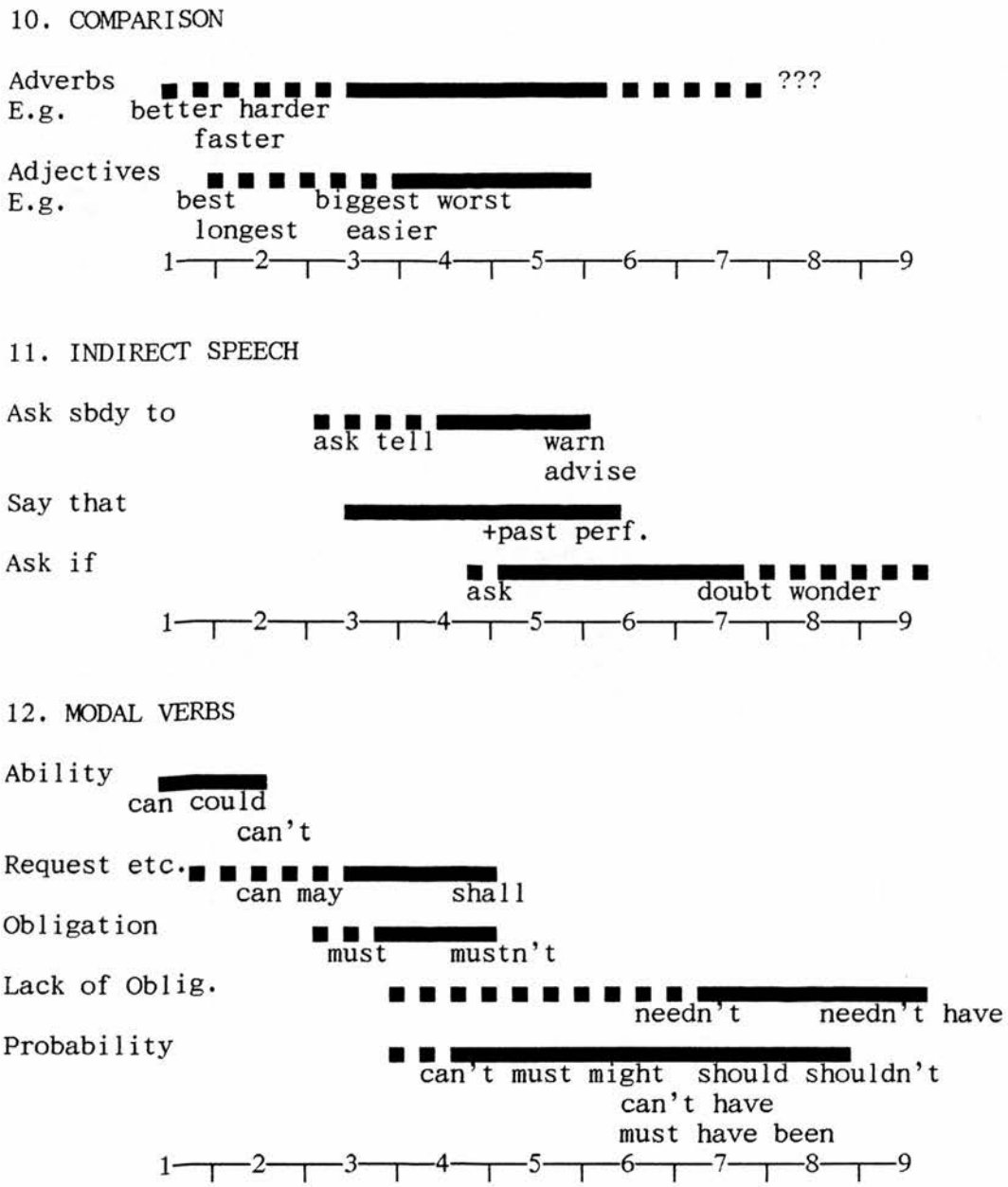


Figure 6.5 Examples of Difficulty envelopes (3)

In particular, they must be seen to represent a summary of one specific, rather small data set, rather than a claim about language proficiency in general. There are simply too few items, trialled on too few persons, to achieve the kind of precision which the figures perhaps appear to lay claim to, or to justify presenting them as having universal significance.

There is evidently some error in the calibration of items, due to sample size and administrative problems at the data collection stage, and also probably to error during estimation (the problem of biassed estimates from extreme scores) and perhaps during linking onto the common bank scale. When one inspects sets of items grouped by language problem, one is generally impressed by the degree to which item difficulty is interpretable; but there are some apparently anomalous cases too, and these make it more difficult to construct difficulty envelopes.

There are gaps in some scales, where items at an appropriate level seem to be missing. The question marks in chart 9, Conditionals, indicate that there are no very difficult items on the Third Conditional, probably because appropriate items were not included in the set.

Even leaving aside the provisional and incomplete nature of these figures, there are important general points to bear in mind concerning their validity.

Firstly, to the extent that they convey some general truth about the development of language proficiency, this is a statistical rather than a psychological truth. Just as no individual family in Great Britain has exactly 2.4 children, so no individual language learner need be expected to conform exactly to this scheme. The item difficulties found in the present study relate to a heterogeneous population of learners. Any conclusions for a developmental theory of language should be drawn with care.

Secondly, however detailed and reliable the characterisation of language proficiency which an item bank like the present one might one day be able to provide, it remains the product of a particular elicitation technique, a particular set of test methods.

After these necessary words of caution, we shall examine two of these difficulty envelopes in more detail, to show how difficulty thresholds were chosen, and how this approach may throw light on the nature of item difficulty.

1) Comparison of Adjectives

Examples of Pre-mastery tasks include:

1. She is the _____ student in the class. (GOOD)

2. My car isn't as fast ____ yours.

In 1) the instruction is to 'complete the sentence with the correct form of the word given in brackets'. The difficulty of items of this type in fact varies widely, depending on the frequency of the word to be supplied. 2) is an example of an item where a function word is to be supplied.

Mastery level is judged to begin with items like:

3. Nobody in the office is fatter than John. (FATTEST)

4. Nobody in the office is fatter than John.
   John ..... in the office.

In 3) the instruction is to write a sentence with the same meaning as the first one, using the given word. Items like 4) are slightly more difficult, apparently because the superlative form of the adjective must additionally be supplied, whereas it is given in the previous example.

Problems of classification begin with items involving transformations between positive and negative, or with switches of subject, such as:

5. The market is less crowded than usual today.
   The market is not ..... .

6. My mother is a better driver than he is. (WORSE)

These seem to be more difficult, and one is tempted to see a method effect in this, given the somewhat unnatural and possibly confusing nature of the task.

There also seems to be a method effect where the word BAD is used as a prompt: it has an undue tendency to elicit the forms BADDER, BADDEST.

Examples of items which are judged to be affected by other more difficult language problems:

7. I've never seen such a bad film.
   That was the .....seen!

8. I can't catch an earlier bus than the 6 o'clock.
   The ..... is at 6 o'clock.

7) is also about the Present Perfect and the never/ever transformation. 8) under-elicits; that is, there are several ways of answering it. Each of these present additional complications. These items thus fall outside the true difficulty range of this language problem.

2) Present Perfect

Examples of Pre-mastery tasks include:

1. He's been working here. (HOW LONG)

2. I ..... this man before now. (NEVER SEE)

1) is a mechanical question-formation task; 2) is also a fairly mechanical task. 2) illustrates a general problem which is particularly acute in the case of items about the Present Perfect: a traditional pedagogic normative rule differs from current usage. Although the item was devised to elicit the Present Perfect, the Past Simple is also perfectly acceptable, at least to most people. The marking scheme has therefore been changed to accept the Past Simple. Pre-mastery here means that a learner *might* use the desired structure, but need not. Many items in the pre-mastery level suffer from this problem.

Mastery level is judged to begin with items like:

3. A: I'm very happy living in Scotland.
   B: And how long .....here?

4. A: Mr Smith is one of our best workers.
   B: And how long .....here?

It is interesting that these items could be either simple or progressive. HOW LONG seems to be the most familiar context for the Present Perfect. Note that the verbs are regular. An item from the higher end of the difficulty range:

5. I'm tired! I ..... ten letters, and still have five more to write. (WRITE)

Generally, less frequent irregular verbs are harder.

Other problems begin to feature with items like these:

6. I have never eaten such a hot curry!
   This is ..... .

7. Porter: Here are your bags.
   Lady: You ..... a mistake. These are not our bags.

8. Newsreader: Here is the news.
   There / be / earthquake / Japan

6) contains a relativization problem, as well as the ever/never transformation. 7) and 8) are the examples of items that rely on conventions of use, or cultural knowledge, which were discussed earlier (5.2.4).

9. This is her first visit to Britain.
   This is the first time she .....

This is a much more difficult item than most in this group; the use of the Present Perfect with 'the first time..' counts as a special, idiomatic case.

An appendix to this chapter provides the full list of items on these two language problems, in order of difficulty, to allow the reader to confirm that the ranking generally agrees with intuition, and also perhaps to pick out the anomalous cases.

The appendix also lists the items on the area of 'making suggestions', to illustrate the point that language problems do not necessarily have to be grammatical. However, if predicting item difficulty is the goal, functional categories will probably not serve very well: the 'Suggestions' list covers the whole range of difficulty, with no obvious 'centre'. This recalls the discussion of whether functional categories provide the best basis for syllabus design (2.2.6). Certainly the list shown in

the appendix points to a longitudinal, 'developmental' dimension to functional categories which is rarely captured in taxonomic, content-defined functional approaches.

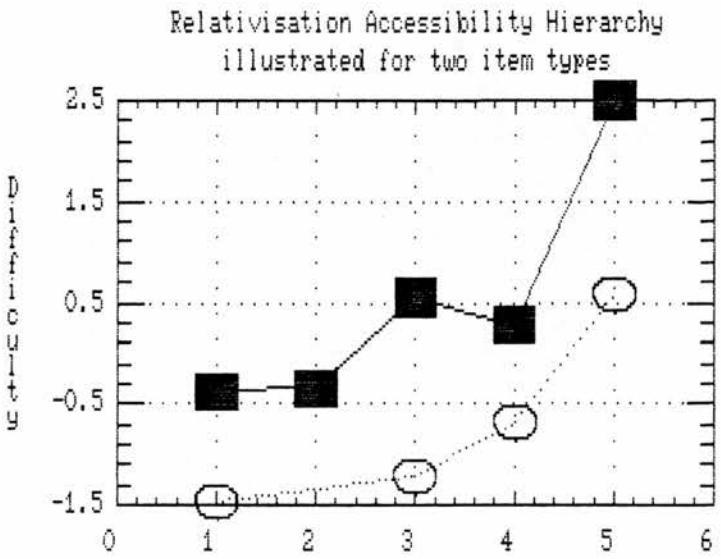
It should be clear that in this chapter we have not attempted to *explain* the inherent difficulty of particular language problems. Rather, the fact of such inherent difficulty is being invoked in order to help explain the difficulty of language test items. This is not begging the question. It seems to be necessary to disentangle these two kinds of difficulty in order to say anything useful about either of them. Certainly, attempts to explain item difficulty by tallying structural features, without explicitly modelling the difficulty of the language problems involved, are unlikely to yield interesting or valid results. It could also be argued that second-language acquisition researchers could benefit from developing a clearer picture of how 'method facets' affect the elicitation of the language data with which they work.

The present study, being very much a study in breadth, is not able to address particular language problems in detail, and the present bank of items is certainly not a test-bed for verifying theories of language difficulty deriving from psycholinguistics or universal grammar. However, let us end this chapter by looking at one such theory, Comrie's accessability hierarchy for relativization (Keenan & Comrie 1977), discussed above (2.2.4), for which the bank contains almost enough data to attempt an analysis.

Figure 6.8 shows five categories of the accessibility hierarchy on the X-axis. They are: (1) subject, (2) direct object, (3) indirect object, (4) object of a preposition, (5) genitive. The theory implies an increasing order of difficulty. Difficulty is plotted on the Y-axis, for two separate item types: a one-word gap fill, where the task was simply to supply the relative pronoun (shown on the graph by the row of circles); and a sentence-completion task, which was predicted to be more

difficult (the row of squares). One data point, the gapfill direct object, is missing; otherwise the values shown are averages of 2 or 3 relevant items.

With one small exception, the plotted points confirm the predictions of the theory. What is perhaps interesting here is that the item type (method) effect and the theoretically predicted effect are both clearly traced, and are thus distinguished from each other.



1 subject
2 direct object
3 indirect object
4 object of a preposition
5 genitive

Top line: difficulty of sentence-completion items
Bottom line: difficulty of one-word gapfill items

Figure 6.8 Comrie's accessibility hierarchy illustrated for two item types

6.6 Appendix: Examples of language problems

1) Comparison of adjectives
2) Present Perfect
3) Functional: Making suggestions

1) Comparison of adjectives

Level

1 / 73

1 My / expensive / yours / is / car / more / than / .

2 / 53

1.5 She is the _____ student in the class. (GOOD)

3 / 980

What's the _____ river in the world? (LONG)

4 / 86

3 My car isn't as fast _____ yours.

5 / 55

That's the _____ building in the town. (BIG)

6 / 991

Keith: Is Tokyo expensive?

Laura: Tokyo? It's ..... expensive city in the world!

7 / 52

3.5 This test is _____ than the last one. (EASY)

8 / 1003

John must be the _____ person in the world! (LAZY)

9 / 204

Nobody in the office is fatter than John. (FATTEST)

10 / 206
I've never seen such a bad film. (WORST)

4 11 / 200
That was the _____ film I've ever seen! (BAD)

12 / 201
Nobody in the office is fatter than John.
John ..... in the office.

13 / 77
Your car isn't as expensive as mine.
My car is ..... .

5 14 / 57
No one in the class is better than Mary. (BEST)

15 / 199
The _____ bus I can catch is at 6 o'clock. (EARLY)

16 / 56
This test isn't as difficult as the last one. (EASIER)

5.5 17 / 203
I've never seen such a bad film.
That was the ..... seen!

6 18 / 202
I can't catch an earlier bus than the 6 'clock.
The ..... is at 6 o'clock.

19 / 260
I have never eaten such a hot curry!
This is ..... .

20 / 54

He's a _____ driver than my mother is. (BAD)

6.5 21 / 205

I can't catch an earlier bus than the 6 o'clock. (EARLIEST)

22 / 198

John is the _____ person in the family. (FAT)

23 / 58

My mother is a better driver than he is. (WORSE)

24 / 257

7.5 The market is less crowded than usual today.

The market is not ..... .

2) **Present Perfect**

Level

1 / 638

2 Have / brilliant / you / seen / film / yet / that / ?

2 / 186

A: Why isn't Brian here today?

B: He ..... to London. (GO)

3 / 982

Dick: ..... ever been to France?

Ian: Me? Yes, many times.

4 / 639

2.5 They / been / abroad / have / never / .

5 / 185

I ..... this man before now. (NEVER SEE)

6 / 984

3 Ann: ..... been learning German?

Jim: Me? Three years.

7 / 170

He's been working here. (HOW LONG)

8 / 172

A: I'm very happy living in Scotland.

B: And how long ..... here?

9 / 190

3.5 Policeman: Do you know this man?

Witness: No, I ..... him before!

10 / 194

I ..... English since I was ten years old. (LEARN)

11 / 91

A: Is this Juan's first visit to Britain?

B: No. He ..... once before.

12 / 961

4 Look at that! You ..... her favourite plate! (BREAK)

13 / 606

A: I usually go to France on the ferry.

B: you / ever / travel / hovercraft ?

14 / 226

4.5 My watch doesn't work. I wonder what ..... to it? (HAPPEN)

15 / 197

A: You speak very good English!

B: I ..... since I was ten.

16 / 196

Man: Hurry up! She ..... for half an hour already!
Woman: Then she can wait a bit longer, can't she?

17 / 75
She / once / been / before / concert / to / has / a / .

18 / 962
A: Is that your car in the drive?
B: Yes, it is.
A: Well, you ..... your lights on. (LEAVE)

19 / 96
A: Are you doing much sightseeing?
B: Yes. We / already / visit / Oxford / Stratford

20 / 93
Oh dear! I ..... horrible mistake!
(MAKE)

21 / 193
Please hurry up. She ..... for half an hour already. (WAIT)

22 / 92
Porter: Here are your bags.
Lady: You ..... mistake. These are not our bags.

23 / 203
I've never seen such a bad film.
That was the ..... seen!

24 / 761
"Has that letter ..... yet?" Mr Johnson asked. (SEND)

25 / 229

6 A: Mary is late.
B: I / wonder / what / happen / her ?

26 / 670
A: Where's Fred?
B: He / go / holiday

27 / 260
I have never eaten such a hot curry!
This is ..... .

28 / 94
A: What's the problem?
B: No problem. I lost my keys but now .....

..... them again. (FIND)

29 / 95
I'm tired! I ..... ten letters, and still have five more to write. (WRITE)

30 / 97
6.5 A: That was a bad fall! Are you OK?
B: Ow! I / afraid / I / break / leg

31 / 723
They've just repaired my car.
I've just had ..... .

32 / 187
7 This isn't the first time she ..... us. She came last year too. (VISIT)

33 / 98
7.5 Newsreader: Here is the news.
There / be / earthquake / Japan

34 / 78

This is only the second time she's been to a concert.
She has only ..... before.

35 / 752

8 My shoes have just been repaired.
I have ..... repaired.

36 / 191

9 This is her first visit to Britain.
This is the first time she ..... .

3) Functional: Making suggestions

Level

1 / 550

2 Why / taxi / a / you / phone / don't / for / ?

2 / 547

Shall / concert / the / we / meet / before / ?

3 / 672

3 You ought to see a doctor. (SHOULD)

4 / 505

3.5 What shall we do this evening?
How _____ going to the cinema?

5 / 292

4.5 My advice to you is to give up smoking.
If I were you ..... smoking.

6 / 650

A: _____ we have a rest?
B: Yes, let's.

7 / 753

She suggested that we stayed at the Sheraton.
"Why ..... at the Sheraton," she said.

8 / 857
"Eat more carrots," said the beautician.
The beautician suggested I .....

9 / 744
5 He advised me to see a doctor.
"If I .....," he said.

10 / 745
5.5 "If I were you I would go to the police," he said.
He advised .....

11 / 89
When shall we go? What _____ Saturday?

12 / 678
6 If / you / coffee / the / some / on / kettle / I'll / put / make / .

13 / 884
You should take up squash. (BETTER)

14 / 894
7.5 If you ..... the seafood menu, I'm sure you would find it delicious. (CHOOSE)

15 / 293
8.5 You should go home at once.
It's high time you .....

16 / 737
9 "Why don't you visit Stratford?" she said.
She suggested ..... Stratford.

7: Conclusion

This study's practical goal is the construction of a testing instrument: an item bank. This has involved three major areas of work: the design and programming of the bank itself, the construction and trialling of items to go in the bank, and the investigation of the nature of the proficiency trait which is depicted by the items, once their difficulty has been found. The final value of the item bank to teachers and learners depends on all of these areas: it needs to be easy to operate - user friendly, that is - and, no less, it needs to produce reliable measures that are of some demonstrable relevance to learners of English in a formal instructional setting.

How user-friendly and attractive the bank will prove to be is something that will become clear only in future, when it becomes available for wider use. The computer-adaptive test, likewise, has yet to be made widely available. But the significance of the present study as an application of Rasch analysis to language testing rests more on the other two areas - the construction and interpretation of the proficiency trait - and this concluding discussion centres on these.

7.1 Rasch analysis and vertical equating

The discussion of item calibration (5.2 above) found that the use of the Rasch model was not without problems. Badly-targetted items with extreme raw scores (i.e. nearly all right or nearly all wrong) were poorly estimated, receiving difficulty values that were biassed away from the mean (i.e. the easy items in a data set were found too easy, the hard items too hard). This bias was evident from inspection of items, and was confirmed by retrialling badly-targetted items on learners of more appropriate level. Retrialled items took on less extreme and altogether more plausible difficulty values.

This effect was especially troublesome (and evident) because of the wide range of difficulty/ability to be fitted to the scale; that is, it is a problem concerning vertical equating of tests. It does not depend on the method of equating used: one-step analysis of a missing data matrix produces comparable results to common-item equating of separately-analysed test forms. It was mitigated by excluding badly-targetted items from analysis, pending retrial. The effect of this is to shorten the scale length - i.e. the range of item difficulty estimates - of each test form, and thus finally to shorten the length of the constructed scale. 6 logits, the effective range of the bank items, is less than is generally reported for scales covering a wide ability range. By setting limits on the data fitted, it appears that the Rasch model can be made to perform satisfactorily, but at a cost in terms of the scale length. This is a loss, because the shorter the scale, the greater the significance of measurement error.

The same bias was evident with the estimation of person abilities. Because of this the transformation tables provided with generated tests, to change raw scores into scale units, have had to be tailored to cut off the top and bottom 20% of possible scores. Abilities estimated from such extreme raw scores are clearly exaggeratedly high or low. Items with raw scores outside the range 20% - 80% are excluded from transformation tables.

This bias in the estimation of person abilities, evident from inspection, has also been empirically confirmed. North (1992) equated the item bank logit scale to Eurocentres' band scale by comparing Item Banker scores with a range of other assessments including teacher impression, a C-test, an oral assessment and a writing test. While this generally worked well, off-target scores (outside the range 20% - 80%) gave ability estimates that were improbable in the light of the other assessments.

There is cause for concern here for anyone interested in using the Rasch model for vertical equating over a large range of ability. The problem identified here deserves further investigation: particularly, we should know just how extreme the percentage-correct score has to be for the effect to become noticeable. The 20% - 80% limit set in the present study was derived from inspection of difficulty and ability estimates, as described above.

7.2 Constructing the trait: model fit

Having worked around the problem of estimating item difficulties, we were able to investigate how well items fitted to a single dimension. The investigation of fit suggested some possible relationships with such factors as:

1) particular criteria for correctness. Changing the marking scheme could make items fit better or worse. It seems that insisting on correct spelling of particular words improves fit, which suggests that spelling accuracy is coherent with the trait as a whole - hence perhaps that the trait relates to learning in a formal instructional setting.

2) item type: there is a suggestion that one-word gapfill items fit worse.

3) item difficulty: we did not find a strong indication that items fitted worse at low (or high) proficiency levels.

4) item content: there is a suggestion that items on grammar points are most coherent with the trait as a whole; items on functional knowledge might fit less well.

5) the influence of L1 transfer. Some cases of probable bias caused by L1 transfer were identified, although this did not seem to be a prevalent problem.

The evidence was clearest in the case of L1 transfer, the other noted relationships lacking statistical significance.

Generally it can be said that the sort of items selected for Item Banker fit readily to a single scale. This is both good and bad news. It is bad news to the extent that misfit analysis frequently fails to identify bad items. Bad items are those which appear to be difficult for reasons extraneous to language use or development. We may hypothesize that the more language-proficient learner, being able to handle whatever linguistic difficulty the item presents, is in a better position to tackle the other conundrums that may be present. Whatever the reason, we found that fit was generally good, and misfit analysis was not hugely revealing of item quality.

7.3 Interpreting the trait

The calibrated items in the bank lie along a coherent dimension. It is, to repeat McNamara's (1990:107) distinction, a *measurement dimension*, not to be confused with 'dimensions of underlying knowledge or ability which may be hypothesized on other, theoretical grounds'. The coherence of the trait suggests that it measures *something*, but what that is requires explanation. How should the trait be named?

The items placed in the bank constitute a heterogeneous, maximally inclusive set, reflecting a weak view of General Language Proficiency as an aggregate sort of measure based on performance in a variety of tasks, rather than the strong view in which GLP reflects a 'real' underlying Unitary Competence. 'General Language Proficiency' might seem to be the best name for the trait. Item Banker tests certainly have much in common with

cloze tests, which are often called tests of GLP. Both are 'indirect', paper-and-pencil tests, lacking 'authentic' communicative purpose, and allowing recourse to explicit, conscious knowledge. But Item Banker items are discrete, and many of them centre on traditional areas of pedagogic grammar. Thus they offer the learner more scope to use (conscious or unconscious) rule-based, grammatical knowledge.

This suggests that the trait might better be named 'Grammatical Competence'. To do so however risks the accusation of 'attempting to generate models of second language acquisition by running theoretically unmotivated data from poorly conceptualized tests through a powerful statistical program', to use Nunan's (1987:156) criticism of the ITESL. That is, it risks being taken as a strong claim about underlying knowledge and abilities. This brings us to the key problem in interpreting the language proficiency trait depicted by the bank: to what extent can it be seen in developmental terms?

The best answer to this might be that it depends on what one means by developmental. We take Swan's (1987:66) view (2.3.2 above) that there is more to language use and development than is captured by studying 'limited data of a very particular kind - ... those phonological and grammatical features which do exhibit variability.' The analysis of item difficulty in Chapter 6 introduced the informal heuristic notion of the *language problem* as a way of grouping items which are in some way about 'the same thing'. A quantitative analysis showed clearly enough that it is mainly the language problem which decides an item's difficulty - a finding which would not surprise any language teacher. The bank trait depicts a broad language proficiency continuum against which the difficulty of a variety of language problems can be mapped. Examples were presented (above, 6.4.2). This mapping does two things: it places each language problem on the scale, showing the range of difficulty it covers; it also *appears* to set different language problems in some relation to each other. This latter aspect should not be over-interpreted as a depiction

of a 'developmental sequence'. It may *describe* a sequence, but it certainly does not *explain* anything, and indeed, there is no reason why many of the language problems identified should be supposed to stand in any functional relationship to each other. The trait as a whole, constructed from a heterogeneous collection of items, should be seen as a matrix in which a number of theoretically interpretable traits may be fixed. It provides the context in which interpretation of subsets of items may be possible.

Thus it is the other aspect of the mapping - the fixing of a 'difficulty envelope' for the language problems identified - which is considered to be significant. Examples were presented (above, 6.4.2) to show how difficulty envelopes can be derived from a qualitative inspection of the items on a given language problem. Inspection of items grouped by language problem shows that the easiness or difficulty of items is largely explicable in terms of factors that belong in a theory of learning. The easiest items on a given language problem typically offer a great deal of support, and thus a simple task. They were termed *pre-mastery* items to indicate that a correct response does not demonstrate a practical mastery of the tested problem. Then come items embodying formulaic use, or use in familiar, 'survival English' functional contexts. Where lexical difficulty is a relevant aspect of a language problem, then commonly-used lexical items make for easiness. Harder items tend to embody use in more abstract, cognitively-demanding contexts, sometimes invoke cultural or conventional knowledge, or an idiomatic special-case usage. These are factors which clearly relate to language use and development. Exactly which features of items are relevant to difficulty depends, of course, on the nature of the problem; this is why such qualitative analysis of items grouped by problem is in many ways more revealing than the quantitative analysis which preceded it.

Inspection also reveals sources of difficulty which are clearly extraneous to the language problem being tested. Items ostensibly about one language problem may be unrepresentatively difficult because they involve other, harder problems, or unrepresentatively easy if there is an acceptable response which represents a 'getout'. Misleading rubrics, garden-path prompts or prompts which elicit the desired response too weakly are all extraneous sources of difficulty. The fixing of criteria for correctness is frequently problematic, given the dichotomous scoring system. Some items turn out to allow several unexpected but unobjectionable answers, and then lose their value as far as the intended language problem is concerned. Sometimes, faced with a gradient of almost-acceptable responses, it seems best to judge correctness narrowly; such items then become perhaps unrepresentatively difficult. Weeding out items which are difficult for extraneous reasons is perfectly feasible, given the discrete-item format, and should be undertaken as better items become available to replace them.

To the extent that item difficulty is explicable in terms of factors which relate to language use and development, then we feel justified in applying the word 'developmental' to the language proficiency trait depicted by the bank items. But reasons for caution have been pointed out (above, 6.4.2) and will be repeated here. The maps of language-problem difficulty presented in Chapter 6 must be taken as statements about the contents of one rather small item bank, rather than universal claims about the English language. There are too few items, and probably there is too much measurement error, to support overmuch generalization at this stage. Secondly, whatever the maps describe is true for a heterogeneous population of learners; it need not be exactly true for any individual. Thirdly, the bank items are accuracy-oriented, and it has been pointed out that accuracy and acquisition are not the same thing (Hakuta 1976, Pienemann 1985).

7.4 Uses of the item bank

The present item bank is intended for use as one aspect of *formative* assessment in a formal instructional setting. It has been argued (above, 2.3.2) that in a teaching setting such a competence-oriented test is in some respects more informative and useful than a 'direct' test of communicative language ability, even if developing communicative language ability is the final goal of the teaching. The detailed information which the bank should make it possible to provide is potentially of benefit to teachers and learners. That at least is the conviction of the present writer, and he counts himself among those English teachers for whom 'integration and communicative language teaching have been a liberating influence in the classroom' (Hamp-Lyons 1989:117).

The appearance of a paper-and-pencil test produced from the bank is traditional and unremarkable. Readers who are sceptical of the utility of discrete-item grammar tests in general may consider that the effort required to build an item bank to produce them is ill-spent. Such readers may be unwilling to accept that to produce tests of known difficulty, which report scores in terms of a single ability scale, is to make a qualitative step forward. They may not notice the difference. Other readers may of course be inclined to err in the opposite direction, placing exaggerated faith in what should be just one aspect of assessment.

The item bank is potentially a very flexible resource. We have contemplated it being used for a variety of purposes (above, 5.1.1): for making classroom exercises, for achievement testing (with a focus on particular language problems) as well as for placement/proficiency testing proper. Various computer-adaptive modes of interaction with the bank are also possible. A 'tutorial' mode in which individual learners could test themselves on chosen content areas, with difficulty automatically adapted to their ability level, seems potentially attractive.

There are clear dangers here, the more so because at present the bank is quite small: one thousand items, spread over all proficiency levels and a variety of content areas, means thin coverage of any single area at a given level. The value of items for proficiency testing is compromised if they have already been used in the classroom for some other purpose. 'In-house' guidelines for use seem necessary: in particular care needs to be taken that proficiency-oriented tests (in contrast to achievement-oriented tests) should include a wide range of content areas.

Achievement-oriented tests also require different interpretation. Where a test focusses on particular language problems which have recently been taught in class, then the 'ability' ratings derived from performance on it will typically be higher (we may predict) than if the test were a less tightly-focussed, proficiency-oriented test. Test users need to be made aware of the difference.

In general, users must be aware of how the scale relates difficulty to ability. The IRT view is that when ability and difficulty coincide (have the same location on the scale) the chance of responding correctly is 50%. Thus a teacher who asks the bank for a proficiency test at the same level as a group of learners can expect those learners to score no more than about 50% on it. From a measurement point of view this is optimum. From a placement point of view, it represents a suitable level of difficulty and challenge for learners to work at: that is, a suitable entry level. But the common-sense view of ability is more in terms of exit level, of the tasks which a learner can perform well on. This perception can fuel the impression that item bank tests are 'too hard'. In practice some accomodation may have to be made to this perception, and guidelines for use should aim to produce slightly higher percentage-correct scores in proficiency tests.

We have insisted that the purpose of the item bank is formative assessment, and that the provision of feedback to learners is an important and legitimate aspect of this. Many readers will see the danger that progress in learning may become too closely identified with scores in item bank tests, and thus that the virtuous feedback circle will degenerate into a vicious circle of 'teaching to the test', with the significance of test scores becoming undermined, and to the general detriment of the teaching programme. Hamp-Lyons (1989), whose criticism of the ITESL has already been mentioned, also draws attention to what she sees as the potentially negative washback effect of that test. She is concerned that by using discrete categories of language structure to define a measurement dimension, ITESL 'makes a statement about how English should be taught (p.117),' and represents a 'backward step ... for language teaching', away, that is, from the sound principles of communication and integration. She asserts that 'language testing ... is a political act (p.111)', and that 'language testing researchers must always be aware of the potential consequences of what they do.'

> Too often, bad tests turn out to be those which have detrimental washback onto the curriculum, which is painfully sensitive to changes in testing practices and very apt to interpret such changes as statements about values and philosophies. (Hamp-Lyons 1989:111)

While we accept this up to a point, we reject the suggestion that a test of structural knowledge is necessarily irrelevant to or in opposition to the goals of a communicatively-oriented course. In the case of the present item bank, we feel that its potential benefits outweigh the dangers of its misuse, particularly given the essentially modest role foreseen for it as a resource for formative assessment at the start of or during a teaching programme. We do not foresee its use for high-stakes testing.

We certainly do not feel that the item bank, to borrow Hamp-Lyon's criticism of the ITESL, 'makes a statement about how English should be taught.' Like any resource, it can be used well or badly. We take the view that most teachers are capable of using it well.

References

Adams, R.J., Griffin, P.E. and Martin, L. 1987. A latent trait method for measuring a dimension in second language proficiency. *Language Testing* 4,1: 8-27

Adjemian, C. 1976. On the nature of interlanguage systems. *Language Learning* 26, 297-320.

Alderson, J.C. 1981a. Reaction to the Morrow Paper. In Alderson J.C. and Hughes A. (eds.).

Alderson, J.C. 1981b. Report of the discussion on general language proficiency. In Alderson J.C. and Hughes A. (eds.)

Alderson, J.C. and Hughes A. (eds.) 1981. *ELT documents 111: Issues in language testing.* The British Council

Allen, J.P.B. and Widdowson, H.G. 1975. Grammar and language teaching. In J.P.B. Allen and S.P. Corder (eds.) *The Edinburgh course in applied linguistics.* Vol. 2. London: Oxford University Press

Allen, P., Cummins J., Mougeon R. and Swain M. 1983. *Development of bilingual proficiency: Second year report.* Toronto, Ont.: The Ontario Institute for Studies in Education

Allen, H.B. and Campbell R.N. 1972. *Teaching English as a Second Language. McGraw Hill*

Allwright, R. 1987. Concluding comments. In Ellis, R. and C. Robert (eds.).

American Psychological Association. 1985. *Standards for educational and psychological testing.* Washington, D.C.: American Psychological Association

Andrich. D. 1988. *Rasch models for measurement.* Newbury Park: Sage Publications

Angoff, W.H. Use of difficulty and discrimination indices for detecting item bias. In Berk, R.A. (ed.) *Handbook of methods for detecting test bias.* Baltimore, MD: The John Hopkins University

Bachman, L. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press

Bachman, L.F. and Palmer A.S. 1982. The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 4:449-65.

Bachman, L.F. and Palmer A.S. 1983. The construct validity of the FSI oral interview. In Oller J.W. (ed.).

Bailey, N., Madden C. and Krashen S. 1974. Is there a 'natural sequence' in adult second language learning. *Language Learning* 24, 235-43.

Bialystok, E. and Sharwood Smith M. 1985. Interlanguage is not a state of mind: An evaluation of the construct for second-language acquisition. *Applied Linguistics* 6.2.

Biggs, J.B. and Collis K.F. 1982. *Evaluating the quality of learning.* New York: Academic Press

Blalock, H. 1969. *Theory construction: from verbal to mathematical formulations.* Englewood Cliffs, NJ: Prentice-Hall

Brown, G. and Yule, G. 1983. *Discourse analysis.* Cambridge: Cambridge University Press

Brumfit, C.J. 1981. Notional Syllabuses revisited: a response. *Applied Linguistics* 2.1

Brumfit, C.J. 1984. *Communicative methodology in language teaching.* Cambridge: Cambridge University Press

Campbell, D.T. and Fiske, D.W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81-105

Campbell, R. and Wales, R. 1970. The study of language acquisition. In Lyons J. (ed.) *New horizons in linguistics.* London: Penguin Books

Canale, M. 1983. On some dimensions of language proficiency. In Oller J.W. (ed.).

Canale, M. and Swain, M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1.1, 1-47.

Carroll, J.B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. Reprinted in Allen and Campbell, 1972.

Carroll, J.B. 1966. The contributions of psychological theory and educational research to the teaching of foreign languages. In Valdman A. (ed.), *Trends in language teaching.* New York: McGraw Hill

Carroll, J.B. 1983. Psychometric theory and language testing. In Oller J.W. (ed.).

Carroll, B.J. and West, R. 1989. *ESU Framework.* London: Longman

Chen, Z. and Henning G. 1985. Linguistic and cultural bias in language proficiency tests. *Language Testing* 2.2

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, Mass.: M.I.T. Press

Choppin, B. 1981. Educational measurement and the item bank model. In *Issues in Evaluation and Accountability.* London: Methuen

Clahsen, H. 1985. Profiling second language development: a procedure for assessing L2 proficiency. In Hyltenstam, K. and Pienemann (eds.)

Cohen, J. and Cohen, P. 1983. *Applied multiple regression/ correlation analysis for the behavioural sciences.* (2nd edition) Hillsdale, N.J.: Erlbaum

Comrie, B. 1984. Why linguists need language acquirers. In Rutherford, W.E. (ed.) *Language universals and second language acquisition.* Amsterdam: John Benjamins

Corder, S.P. 1967. The significance of learners' errors. *International Review of Applied Linguistics* 4, 161-9.

Corder, S.P. 1981. *Error analysis and interlanguage.* Oxford: Oxford University Press

Criper, C. 1981. Reaction to the Carroll paper (2). In Alderson J.C. and Hughes A. (eds.)

Cummins, J. 1980. The cross-lingual dimensions of language proficiency: implications for bilingual education and the optimal age question. *TESOL Quarterly* 14:175-187

Cummins, J. 1983. Language proficiency and academic achievement. In Oller J.W. (ed.).

Curtis, M.E. 1987. Cognitive analyses of verbal aptitude tests. In Freedle, R.O. and Duran, R.P (eds.)

Cziko, G. 1984. Some problems with empirically-based models of communicative competence. *Applied Linguistics* 5.1, 23-38.

Davies, A. 1978. Language testing. *Language Teaching and Linguistics Abstracts* 11:145-159, 215-231.

Davies, A. 1981. Reaction to the Palmer and Bachman and the Vollmer papers. In Alderson J.C. and Hughes A. (eds.).

Dublin, F. and Olshtain E. 1986. *Course design.* Cambridge: Cambridge University Press

Dulay, H.C. and Burt M.K. 1972. Goofing: an indication of children's second language learning strategies. *Language Learning* 22, 235-52.

Dulay, H.C. and Burt M.K. 1973. Should we teach children syntax?. *Language learning* 23, 235-52.

Dulay, H.C. and Burt M.K. 1974. Natural sequences in child second language acquisition. *TESOL Quarterly* 8, 129-36.

Duran, R.P., Canale M., Penfield J., Stansfield, C.W. and Liskin-Gasparro, J.E. 1987. TOEFL from a communicative viewpoint on language proficiency: a working paper. In Freedle, R.O. and Duran, R.P (eds.)

Ebel, R.L. 1979. *Essentials of Educational Measurement.* Englewood Cliffs, New Jersey: Prentice-Hall

Ellis, R. 1985. *Understanding Second Language Acquisition.* Oxford: Oxford University Press

Ellis, R. and Roberts, C. 1987. Two approaches for investigating second language acquisition in context. in Ellis, R. and C. Robert
(eds.).

Ellis, R. and Roberts, C. (eds.) 1987. *Second Language Acquisition in Context.* Englewood Cliffs, New Jersey: Prentice-Hall

Faerch, C. and Kasper, G. 1986. Cognitive dimensions of language transfer. In Kellerman, E. and M. Sharwood Smith (eds).

Farhady, H. 1983a. On the plausibility of the unitary language proficiency factor. In Oller J.W. (ed.).

Farhady, H. 1983b. The disjunctive fallacy. In Oller J.W. (ed.).

Farhady, H. 1983c. New directions for ESL proficiency testing. In Oller J.W. (ed.).

Felix, S.W. 1978. *Linguistische Untersuchungen zum naturlichen Zweitsprachenerwerb.* Munchen: W. Fink

Felix, S.W. 1982. *Psycholinguistische Aspekte des Zweitsprachernerwerbs.* Tubingen: Narr

Fellbaum, C. 1987. A preliminary analysis of cognitive-linguistic aspects of sentence completion tasks. In Freedle, R.O. and Duran, R.P (eds.)

Freedle, R.O. and Duran, R.P. (eds.). 1987. *Cognitive and Linguistic Analyses of Test Performance.* Norwood, New Jersey: Ablex Publishing Corporation

Freedle, R.O. and Fellbaum C. 1987. An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In Freedle, R.O. and Duran, R.P (eds.)

Goldstein, H. 1981. Limitations of the Rasch model for educational assessment. In *Issues in Evaluation and Accountability.* London: Methuen

Gregg, K.R. 1984. Krashen's Monitor and Occam's Razor. *Applied Linguistics* 5.2, 79-100.

Griffin, P.E., Adams, R.J., Martin, L. and Tomlinson, B. 1986. *Proficiency in English as a second language. The development of an interview test for adult migrants.* Melbourne: Ministry of Education (Schools Division), Victoria

Griffin, P.E., Adams, R.J., Martin, L. and Tomlinson, B. 1988. An algorithmic approach to prescriptive assessment in English as a Second Language. *Language Testing* 5,1: 1-18

Guttman, L. 1950. The basis for scalogram analysis. In Stouffer. S.A. (ed.) *Measurement and Prediction.* New York: Wiley

Hahn, A. 1982. Fremdprachenunterricht und Spracherwerb. Ph.D. Dissertation University of Passau.

Hakuta, K. 1976. Becoming bilingual: a case study of a Japanese child learning English. *Language Learning* 26, 321-51.

Hakuta, K. and Cancino, H. 1977. Trends in second-language acquisition research. *Harvard Educational Review* 47, 294-316.

Halliday, M.A.K. 1973. *Explorations in the functions of language.* London: Edward Arnold

Hambleton, R.K. and Swaminathan, H. 1985. *Item Response Theory - Principles and applications.* Boston: Kluwer-Nijhoff

Hamp-Lyons, L. 1989. Applying the partial credit method of Rasch analysis: language testing and accountability. *Language Testing* 6,1: 109-118

Henning, G. 1984. Advantages of latent trait measurement in language testing. *Language Testing* 1,2, 123-133

Henning, G. 1987. *A Guide to Language Testing Development Evaluation Research.* Cambridge, Mass.: Newbury House

Henning, G., Hudson, T. and Turner, J. 1985. Item response theory and the assumption of unidimensionality for language tests. *Language Testing* 2.2: 141-154

Higgs, T.V. and Clifford, R.T. 1982. The push towards communication. In Higgs, T.V. (ed.) *Curriculum, competence and the foreign language teacher.* Skokie, IL.: National Textbook Co.

Hughes, A. 1981. Reaction to the Palmer and Bachman and the Vollmer papers. In Alderson J.C. and Hughes A. (eds.).

Hughes, A. and Porter, D. (eds.) 1983. *Current developments in language testing.* London: Academic Press

Hulstijn, J.H. 1985. Testing second language proficiency with direct procedures. A comment on Ingram. In Hyltenstam, K. and Pienemann (eds.)

Hulstijn, J.H. 1985. Second language proficiency: an interactive approach. In Hyltenstam, K. and Pienemann (eds.).

Hyltenstam, K and Pienemann M. (eds.) 1985. *Modelling and assessing second language acquisition.* Clevedon, Avon: Multilingual Matters

Hymes, D. 1967. Models of the interaction of language and social setting. In J. Macnamara (ed.) *Problems of bilingualism.* ?.

Hymes, D. 1972. On communicative competence. In Pride J.B. and Holmes, J. (eds), *Sociolinguistics.* Harmondsworth: Penguin Books

Ingram, D.E. 1985. Assessing proficiency: an overview on some aspects of testing. In Hyltenstam, K. and Pienemann (eds.).

Ingram, E. 1978. The psycholinguistic basis. In Spolsky, B. (ed.) *Advances in language testing: Series 2, Approaches to language testing.* Arlington, Virginia: Center for Applied Linguistics

Johnson, K. 1982. *Communicative syllabus design and methodology.* Oxford: Pergamon Press

Jordens, P. 1986. Production rules in interlanguage: evidence from case errors in L2 German. In Kellerman, E. and M. Sharwood Smith (eds).

Kay, P. 1987. Three properties of the ideal reader. In Freedle, R.O. and Duran, R.P (eds.)

Kean, M. 1986. Core issues in transfer. In Kellerman, E. and M. Sharwood Smith (eds).

Keenan, E. and Comrie B. 1977. Noun phrase accesibility and universal grammar. *Linguistic Inquiry* 8, 63-99.

Kellerman, E. 1979. The problem with difficulty. *Interlanguage Studies Bulletin* 4, 27-48.

Kellerman, E. and Sharwood Smith M. 1986. *Crosslinguistic Influence in Second Language Acquisition.* Oxford: Pergamon Press

Kohn, K. 1986. The analysis of transfer. In Kellerman, E. and M. Sharwood Smith (eds).

Krashen, S. 1981. *Second language acquisition and second language learning.* Oxford: Pergamon Press

Krashen, S. 1982. *Principles and practices of second language acquisition.* Oxford: Pergamon Press

Krashen, S. 1985. *The input hypothesis: issues and implications.* London: Longman

Labov, W. 1966. *The social stratification of English in New York City.* Washington, DC: Center for Applied Linguistics

Lado, R. 1961. *Language testing: the construction and use of foreign language tests.* New York: McGraw-Hill Book Company

Lamotte, J., Pearson-Joseph, D. and Zupko, K. 1982. A cross-linguistic study of the relationships between negation stages and the acquisition of noun-phrase morphology. Term paper, Ed. 676, University of Pennsylvania.

Larsen-Freeman, D. 1975. The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly* 9, 409-14.

Lee, O.K. 1991. Convergence: Statistics or substance? *Transactions of the Rasch Measurement SIG, American Educational Research Association* 5, 3.

Lee, Y.P., Fok A., Lord R. and Low, G. (Eds) 1985. *New Directions in Language Testing.* Oxford: Pergamon Press

Lenneberg, E.H. 1967. *Biiological foundations of language.* New York: Wiley

Lightbown, P.M. 1985. Can language acquisition be altered by instruction? In Hyltenstam, K. and Pienemann (eds.).

Linacre, J.M. 1989. *Many-faceted Rasch meaurement.* Chicago: MESA Press

Loevinger, J. 1954. The attenuation paradox in test theory. *Psychological Bulletin* 51: 493-504

Long, M. 1985. A role for instruction in second language acquisition: task-based language teaching. In Hyltenstam, K. and Pienemann (eds.).

Lord, F.M. 1980. Some how and which for practical tailored testing. In Van der Kamp *et al* (eds.).

Lord, F.M. and Novick M.R. 1968. *Statistical theories of mental test scores.* New York: Addison-Wesley

Lowe, P. Jr. 1988. The unassimilated history. In Lowe, P. and Stansfield (eds) *Second Language proficiency assessment: Current issues.* Englewood Cliffs, NJ: Prentice-Hall

Lowe, P. Jr. 1982. *ILR Handbook on oral interview testing.* Washington DC: DLI/LS Oral Interview Project

Lumsden, J. 1976. Test theory. In Rosenzweig, M.R. and Porter, L.W. (eds.), *Annual review of psychology.* Palo Alto, CA: Annual Reviews Inc.

McLaughlin, B. 1987. *Theories of second-language learning.* London: Edward Arnold

McNamara, T.F. 1990 *Assessing the second language proficiency of health professionals.* Unpublished Ph.D. dissertation, University of Melbourne

Milanovic, M. 1988. The construction and validation of a performance-based battery of English language progress tests. Unpublished Ph.D. dissertation, University of London.

Morrow, K. 1981. Communicative language testing: revolution or evolution. In Alderson J.C. and Hughes A. (eds.).

Munby, J. 1978. *Communicative syllabus design.* Cambridge: Cambridge University Press

Neisser, U. 1967. *Cognitive Psychology.* New York: Appleton, Century, Crofts

North, B. 1992. Item Banker calibration study. MS.

Nunan, D. 1987 Methodological issues in research. In Nunan, D. (ed.) *Applying second language acquisition research.* Adelaide, SA: National Curriculum Resource Centre

Oller, J.W. Jr. 1974. Expectancy for successive elements: key ingredient to language use. *Foreign Language Annals* 7, 105-18.

Oller, J.W. Jr. 1978. How important is language proficiency to IQ and other educational tests?. Oller J.W. Jr and Perkins K. (eds) *Language in education: testing the tests.*Rowley, Mass.: Newbury House

Oller, J.W. Jr. 1983. A general language proficiency factor. In Oller J.W. (ed.).

Oller, J.W. (ed.) 1983. *Issues in Language Testing Research.* Rowley, Mass.: Newbury House

Perkins, K. and Linnville, S.E. 1987. A construct definition study of a standardized ESL vocabulary test. *Language Testing* 4,2:125-141

Pica, T. 1982. The role of language context in second language acquisition. Review article MS.

Pica, T. 1985. Linguistic simplicity and learnability: implications for language syllabus deisgn. In Hyltenstam, K. and Pienemann (eds.).

Pienemann, M. 1985. Learnability and syllabus construction. In Hyltenstam, K. and Pienemann (eds.).

Pollitt, A. 1990. Diagnostic Assessment through item banking. In Entwistle, N. (ed.) *Handbook of educational ideas and practices.* London: Croom Helm

Pollitt, A. and Hutchinson, C. 1986. The validity of reading comprehension tests: What makes questions difficult?. In D. Vincent, A.K. Pugh and G. Brooks (eds.) *Assessing Reading.* London: Macmillan

Pollitt, A. and Taylor, L. 1991. Question level bias in cloze questions - an L1 transfer effect. Paper presented at the First European Language Testing Symposium. Jyväskylä

Popham, W.J. 1978. *Criterion-referenced measurement.* Englewood Cliffs, N.J.: Prentice-Hall

Porter, J. 1977. A cross-sectional study of morpheme acquisition in first-language learners. *Language Learning* 27, 47-62.

Porter, D. 1983. The effect of quantity of context on the ability to make linguistic predictions: a flaw in a measure of general proficiency. In Hughes, A. and Porter, D. (eds.)

Prabhu, N.S. 1987. *Second language pedagogy.* Oxford: Oxford University Press

Rasch, G. 1960. [1980]. Probabilistic Models for some Intelligence and Attainment Tests. Expanded edition with a foreword and an afterword by B.D. Wright. Chicago: The University of Chicago Press

Rutherford, W.E. 1987. *Second language grammar: learning and teaching.* London. Longman

Sang, F., Schmitz B., Vollmer H.J., Baumert J. and Roeder P.M. 1986. Models of second language competence: a structural equation approach. *Language Testing* 3, 1, 54-79.

Schachter, J. 1974. An error in error analysis. *Language Learning* 24, 205-14.

Schachter, J. and Celce-Murcia, M. 1979. Some reservations concerning error analysis. *TESOL Quarterly* 11, 441-51.

Selinker, L. 1969. Language transfer. *General Linguistics* 9.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics* 10, 209-31.

Skehan, P 1988-89. State of the art article: Language testing. Parts 1 and 2. *Language Teaching* 21.4, 22, 1. Cambridge: Cambridge University Press

Smith, R.M. 1991. *Assessing unidimensionality for Rasch measurement.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago

Sorace, A. 1985. Metalinguistic knowledge and language use in acquisition-poor environments. *Applied Linguistics* 6.1: 239-254

Spolsky, B. 1973. What does it mean to know a language? Or, how do you get someone to perform his competence?. In Oller, J.W. Jr and Richard (eds.) *Focus on the learner: Pragmatic Perspectives for the Language Teacher.* Rowley: Newbury House

Spolsky, B. 1975. Language testing: art or science?. Paper presented at the Fourth AILA International Congress, Stuttgart.

Spolsky, B. 1988. Test review: P.E. Griffin *et al.* 1986, Proficiency in English as a second language. (1) The development of an interview test for adult migrants. (2) The administration and creation of a test. (3) An interview test of English as a second language. *Language Testing* 5,1: 120-124

Stenner, Smith and Burdick, A. Jackson 1983. Toward a theory of construct definition. *Journal of Educational Measurement* 20.4.

Stevenson, D.K. 1981. Beyond faith and face validity: the multitrait-multimethod matrix and the convergnt and discriminant validity of oral proficiency tests. In Palmer, A.S., Groot, P.J.M. and Trosper, G.A. (eds). *The Construct Validation of Tests of Communicative Competence.* Washington DC: TESOL

Swan, M. 1987. Non-systematic variability: a self-inflicted conundrum?. In Ellis, R. and C. Robert (eds.).

Tall, G. 1981. The possible dangers of applying the Rasch model to school examinations and standardized tests. In *Issues in Education and Accountability.* Methuen London

Tarone, E. 1983. On the variability of interlanguage systems. *Applied Linguistics* 4. 142-63.

Tarone, E. 1987. Methodologies for studying Variability in second language acquisition. In Ellis, R. and C. Robert (eds.).

Thurstone, L.L. 1959. *The Measurement of Values.* Chicago: University of Chicago Press

Traub, R.E. 1983. A priori considerations in choosing an item response model. In Hambleton, R.K. (ed.) *Applications of item resonse theory.* Vancounver, BC: Educational Research Institute of British Columbia

Van der Kamp, L.J.Th., W.F. Langerak and D.N.M. de Gruijter (eds.) 1980. *Psychometrics for Educational Debates.* New York: Wiley

Van Ek, J.A. 1976. *Significance of the threshold level in the early teaching of modern languages.* Strasbourg: Council of Europe

Vollmer, H.J. 1981. Why are we interested in general language proficiency?. In Alderson J.C. and Hughes A. (eds.).

Vollmer, H.J. and Sang, F. 1983. Competing hypotheses about second language ability: a plea for caution. In Oller J.W. (ed.).

Weir, C.J. 1981. Reaction to the Morrow paper. In Alderson J.C. and Hughes A. (eds.)

Weir, C.J. 1988. Construct validity. In Hughes, A., Porter, D. and Weir, C. (eds.) *ELTS Validation Project; proceedings of a conference held to consider the ELTS Validation Project Report. English Language Testing Service Report* 1 (ii). London: British Council / University of Cambridge Local Examinations Syndicate

Wilkins, D.A. 1981. Notional Syllabuses revisited. *Applied linguistics* 2.1.

Wise, S.L. 1989. Research on the Effects of Administering Tests via Computers. *Educational Measurement Issues and Practice* 8.3.

Wode , H. 1981. *Learning a second language. Vol. 1, An integrated view of language acquisition.* Tubingen: Narr

Wood, R. 1978. Fitting the Rasch model - a heady tale. *British Journal of Mathematical and Statistical Psychology* 31: 27-32.

Wright, B.D. 1988. Georg Rasch and measurement. *Transactions of the Rasch Measurement SIG, American Educational Research Association* 2, 3.

Wright, B.D. and Stone M.H. 1979. *Best Test Design.* Chicago: MESA Press

Wright, B.D. and Masters, G.N. *Rating scale analysis.* Chicago: MESA Press